SOFTEN THE PAIN, INCREASE THE GAIN: INTERVENTIONS THAT MITIGATE THE INFLUENCES OF NEGATIVE FEEDBACK AND IMPROVE ONLINE FEEDBACK EXCHANGE

BY

YU WU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Professor Brian P. Bailey, Chair
Professor Karrie Karahalios
Associate Professor Ranjitha Kumar
Dr. Elizabeth Churchill, Google LLC

ABSTRACT

Content creators can experience negative feedback when sharing work on online platforms. In this thesis, we define feedback written in an unnecessarily harsh tone as negative feedback. Content creators who receive negative feedback report lower levels of affective states and generate lower quality work. In my thesis, I quantify the influences of negative feedback and report results from three experiments testing novel techniques that aimed at mitigating these influences: self-directed coping activities, valence-based ordering of feedback, and empathy arousal. In the first experiment, we investigate the efficacy of three coping activities: self-affirmation, expressive writing, and distraction. Participants (N=480) revised their essays after performing a coping activity. We find expressive writing encourages essay revision, distraction improves affective states and perception of feedback provider, and self-affirmation has no statistical effects on the outcome measures. In the second experiment, we present feedback in a valence-based order. Participants (N=270) write a story and revise it after receiving feedback in different valence orders. Our main result is that presenting negative feedback last improves content creators' affective states and perception of the feedback. In the third experiment, we explore ways to discourage users from generating negative feedback. Participants (N=205) read a narrative about the content creators before providing feedback. We also explore how an ingroup framing in task instructions mediates the effectiveness of narrative empathy. Our results show both narrative empathy and ingroup framing increases feedback providers' invested effort and the quality of the feedback. The techniques investigated in these experiments are situated within a broader design space for feedback exchange. We hope these techniques promote the generation of more constructive and considerate feedback in online platforms, thereby helping content creators improve their work and benefit from the feedback exchange process.

To my parents, for always striving to give me the best education possible.

ACKNOWLEDGMENTS

My doctorate training has been a long journey. I couldn't have possibly finished it without the generous assistance from many people. Here, at the beginning of my thesis, I'd like to express my sincere gratitude to the people who have helped me in the past seven years.

First, I'd like to thank my advisor, Professor Brian P. Bailey, for his constant support and patient mentoring over the years. As someone who has no prior Human-Computer Interaction (HCI) research experience (not even one class taken!) before my Ph.D., I deeply appreciate Brian's belief in my potentials. Although looking back now, the five papers we co-authored may indicate picking me might not be such a risky bet, all these contributions won't be possible without Brian's patient coaching. I still vividly remember the days and nights we spent together working on my very first ACM CHI Conference on Human Factors in Computing Systems paper submission, during which you showed me how to report world-class research by revising my entire draft word by word. Also, thank you for always holding a high bar for any work I did, no matter if it's to be published or not. And thank you for raising the bar slightly higher every time I reached it. You have pushed me to a height that I didn't know I could have reached.

I also want to thank the other three distinguished researchers that served on my committee, Prof. Karrie Karahalios, Prof. Ranjitha Kumar, and Dr. Elizabeth Churchill. Karrie, although you are not my official advisor, I'd like to thank you for everything I learned from you during our research collaborations and when I worked as a teaching assistant for your course. Also, thank you for looking out for the HCI Ph.D. students and finding us a dedicated lab. Ranjitha, thank you for your advice on thesis topic choice early in my Ph.D. And thank you for offering an AI-focused perspective throughout the process and making my contributions more practical as a result. Elizabeth, thank you for your kind guidance both during my internship at Google and at a later stage of my Ph.D. As someone who eventually decides to join the industry as a product researcher, I feel very fortunate that my industry career started at your team and with you as the perfect role model.

I'd also like to express my gratitude to the staff members at the Thomas M. Siebel Center, especially Elaine Wilson, Sherry Unkraut, Maggie Metzger Chappell, Madeleine Garvey, and Viveka Kudaligama. Thank you for your patient assistance and all those kind reminders when I failed to get back to you promptly! As one of the largest public universities in the U.S., sometimes it can be tricky to figure out and follow the right administrative processes. Thank you for making it as easy as possible for me to navigate through these.

And of course, I'd like to thank all my labmates, including Grace Yen, Sneha Kumaran, Patrick Crain, Emily Hastings, Roshanak Zilouchian, Anbang Xu, Robert Deloatch, Hidy Kong, Helen Wauck, Kristen Vaccaro, Eric Yen, Yi-Chieh Lee, Sanorita Dey, John Lee, Motahhare Eslami, Jennifer Kim, Nikita Spirin, Vera Liao, Mary Pietrowicz, Amy Bretches, Ziang Xiao, Mingkun Yang, Pingjing Yang, Shiliang Zuo, Gina Do, Sebastian Rodriguez, Doris Lee, Jinda Han, Joon Park, Silas Hsu, Ti-Chung Cheng, Wendy Shi, Farnaz Jahanbakhsh. You all are my favorite people! Without your company, I couldn't imagine how miserable my social life could have been. Catching paper deadlines together, going to conferences together, all those HCI social events, the game nights, the group lunches/dinners, these all are among my most fond memories during my Ph.D. I feel very fortunate to get to know you all.

I also like to thank Liqi Xu, my partner, and a fellow Ph.D. student at UIUC. Thanks to you, the past five years have not only been the most intellectually stimulating in my life but also the happiest five years in my life. Thank you for all the smart, funny, encouraging, and warm conversations. With you by my side, I wouldn't mind doing the whole thing all over again for another dozen times.

At last, I'd like to thank my parents. As a middle-class family in a second-tier city in China, we usually don't have lots of money to splurge. But you always, always, have made my education the very highest priority. Growing up, I might not have the shiniest toys compared to my friends, but there had never been a time when you complained about buying me books and paying for my after-school classes. No matter how expensive they were. And again, thank you for supporting my decision to come to the U.S. for undergrad education. Being the unworldly high-schooler back then, I didn't understand how much financial strain this would put on you two and how much harder you two needed to work for the next decade. You are the best parents that any child could ever wish for.

Thank you all. Without your help, nothing I've accomplished is possible.

TABLE OF CONTENTS

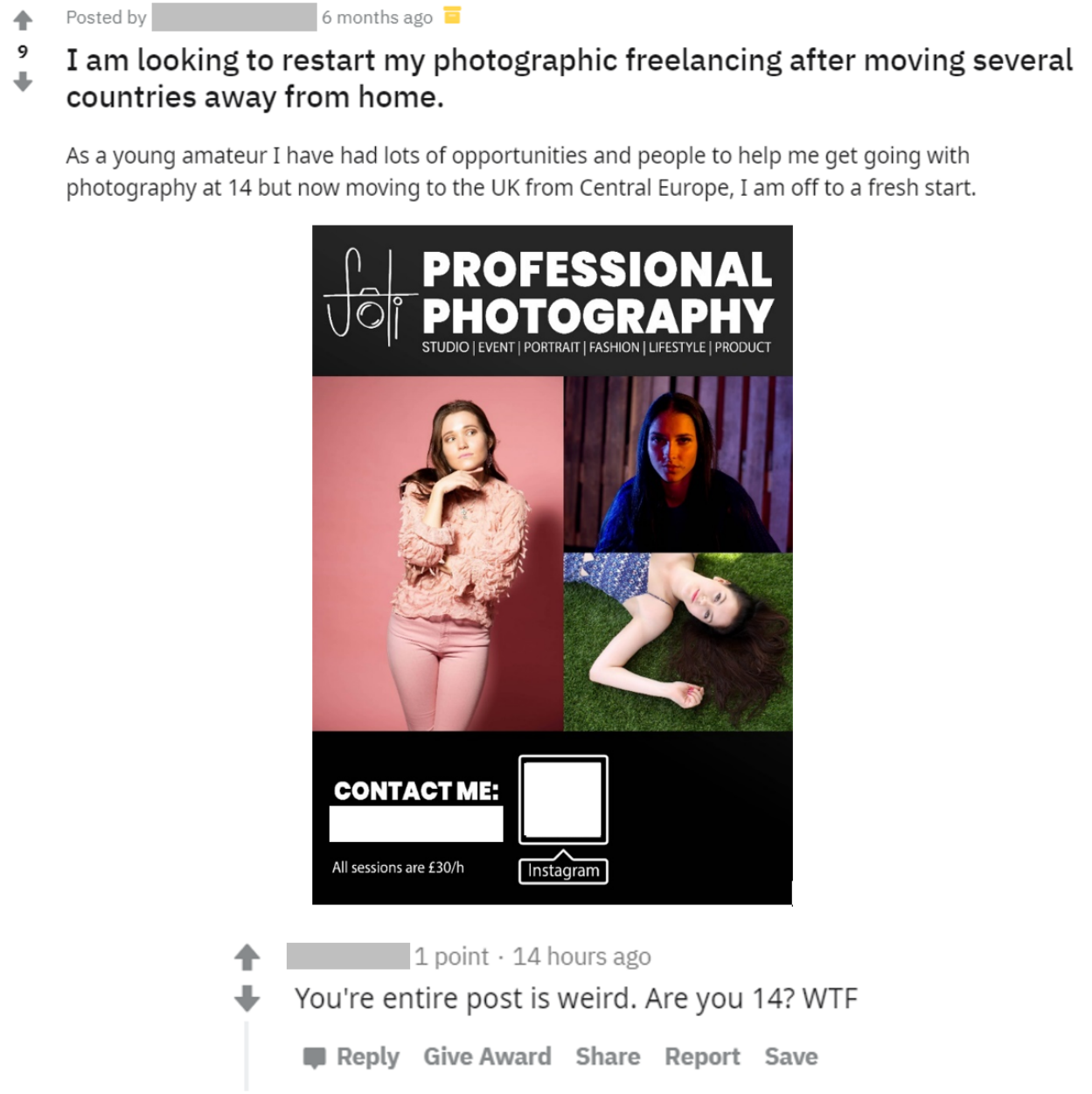
CHAPTER 1: INTRODUCTION

Online feedback collection platforms enable content creators to amass a wide range of critiques quickly [1]. Such platforms are particularly useful for amateurs and novice content creators who do not have extensive professional networks to seek feedback from. A common problem is content creators receiving feedback with negative valence, which is a comment — directed at either the work or the content creator themselves — written using an unnecessarily negative tone (See Figure 1.1). We refer to this kind of feedback as negative feedback. For example, when a novice content creator posts a web page design for feedback on a popular online forum, one of the feedback providers wrote, "the design is so bad that no one wants to criti[que it]." This problem is common on feedback collection platforms for various creative design types, including graphic design, essay writing, web design, etc. This problem exists in face-to-face settings but becomes more extensive online for two reasons. First, the mask of anonymity allows more aggressive behavior on online platforms [2]. Second, anti-social users are more active than average users [3, 4]. Moreover, negative feedback snowballs, since exposure to negative feedback encourages users to generate more such content [5]. Exposure to negative feedback encourages users to generate more such content [5]. Prior work shows negative valence information substantially reduces people's affective states [6] and erodes their task performance [7, 8]. Also, the negativity masks constructive advice in the feedback and dissuades content creators from making effective revisions. For these reasons, we envision a future where feedback exchange platforms utilize various mechanisms to encourage constructive feedback generation and equip content creators with strategies and techniques to minimize the unfavorable influences of receiving negative feedback.

1.1 EXISTING APPROACHES

Negative feedback is common online beyond feedback collection communities. It is not difficult to find aggressive content when people discuss various socio-political topics. Platform designers and researchers have taken different approaches to address this problem. To examine existing approaches in a systematic fashion, here I propose a two-dimensional design space to include existing solutions. The first dimension is who manages the process that inhibits negative content generation. Within this dimension, the four key roles are content creators, content receivers, administrators, and automated mechanisms on platforms. The other dimension is at which step of the content generation process the approach controls negative content. There are five steps determining: who provides content, how content is

Figure 1.1: A content creator received the feedback below when seeking feedback on this design advertising a photography service. The platform clearly stated the target audience is amateur designers and encourage constructive feedback. It also employs human moderators and rule-based algorithms to regulate community interactions. However, as this post illustrated, content creators still receive negative feedback at times. The design and the feedback was collected from Reddit.com in 02/2020. Personal contact information is masked to protect privacy.



composed, who receives content, how content is presented, and how content is consumed. This 4x5 design space (see Table 1.1) describes existing approaches using a generic formula: which key role enables an intervention at which step of the process to control negative content. For example, a community forum may rely on feedback providers to proactively report malicious users in order to deter negative feedback by controlling who provides feedback; platform administrators may filter contents based on user traits (e.g., age, expertise, personal preference, etc.) to control who receives content; automated mechanisms may use an algorithm to aggregate the feedback and mask the negative content to control how feedback is presented.

	Feedback Provider	Feedback Receiver	Administrator	Automated Mechanism
Step 1: Feedback Provider Selection	Reporting Malicious Users [9, 10]	User-Specific Whitelist / Blacklist [11]	Platform-Wide Whitelist / Blacklist [11]	Learning-Based Detection and Banning [12]
Step 2: Feedback Composition	Rubrics [13], Model Feedback [14]	**Narrative Empathy**, Feedback Type Specification [15]	Topic-Specific / Ad Hoc Moderation [16]	**Ingroup Framing**, Feedback Scaffolding [15, 17]
Step 3: Feedback Recipient Selection	Distributed Moderation [9, 10], User-Trait Based Filtering [18]	Self-Selected Filtering [19]	User-Trait Based Filtering [18]	Algorithmic Content Ranking [20], Learning-Based Moderation [21]
Step 4: Feedback Presentation	Peer Popularity Voting [22]	Learning-Based Feedback Selection [23, 24]	Ad Hoc Censoring [25]	**Valence-Based Order**, Feedback Aggregation [26], Positive Framing [27]
Step 5: Feedback Consumption	Content Tagging [28]	**Coping Activities**	Content Tagging [28]	**Coping Activities**

Table 1.1: The proposed design space includes most existing approaches to control negative feedback. The Related Work section below will unpack each approach in detail. The approaches proposed in my thesis are in bold.

Generally, existing approaches follow the same philosophy: shielding users from consuming negative feedback. This philosophy has two shortcomings. First, it adopts a strict binary classification of all feedback, labeling each piece as either negative or non-negative instead of treating feedback negativity as a spectrum. Based on the labeling, users either receive the feedback or not. No middle ground exists, and both extremes are undesirable. Censoring negative content in broad strokes may render the platform ineffective or even dysfunctional; being too lenient may impair the user experience of receiving feedback. Second, community guidelines and instructions rely on the initiative of the feedback providers. Many interventions become ineffective if users do not have strong motivations to help the content creators. They may be reluctant to follow instructions, and some may ignore interventions altogether. My proposed work fills this void, and we describe how our approaches address these shortcomings in the following section.

1.2 PROPOSED SOLUTIONS

In my thesis, I developed and tested three novel techniques that complement existing approaches to mitigate the impact and generation of negative feedback, without censoring or inhibiting the exchange of constructive feedback. Specifically, I proposed a set of recipient centered coping interventions and an algorithmic mechanism that ease the consumption of negative feedback. I also devised two techniques that facilitate content generation and encourage feedback providers to be more supportive and produce higher quality content.

In Chapter 4, we examine three theory-based coping activities: expressive writing, distraction, self-affirmation. Participants (N=480) receive feedback sets with different balances of neutral and negative valence content and revise their essays after performing the assigned activity. We measure participants' affective states, extents of revision, and their perceptions of the feedback and its providers. We select these three activities because they correspond to three general approaches people take in the face of ego-threatening information: fight, flight, or self-affirm [29]. Prior work demonstrates these three activities have notable potentials for increasing users' resilience to negative valence information [30, 31, 32, 33, 34, 35, 36]. To evaluate and compare these three coping activities, we conduct an online experiment where participants write an essay on a complex social issue and revise their essay based on a provided feedback set. We measure participants' affective states and the extent of revisions to quantify the impact of negative feedback and coping activities.

In Chapter 5, we explore whether more subtle changes to the existing feedback collection process, such as presenting feedback in different orders, can alleviate this problem. Specifically, we examine whether positive valence feedback could mitigate the influence of the

negative valence feedback in the collection. Besides the valence order factor, we include a feedback source factor to analyze how it affects participant perceptions of the feedback and interacts with the effects of manipulating the order of the feedback in the collection. We conduct an online experiment where participants (N=270) write a children's story and later revise it based on a set of feedback. The task interface presents the feedback in different valence orders together with source identity cues based on the experiment condition. To evaluate the proposed mechanism, we measure participants' affective states, perceptions of the feedback and its source, revision extent, and story quality.

In Chapter 6, we examine the effectiveness of empathy arousal in inducing positive valence feedback generation. Prior work shows a high level of empathetic feelings encourages users to help people in distress [37]. Within the context of online feedback collection, the benefits of empathy may translate into encouraging more considerate and more helpful feedback (prosocial behavior) and a stronger stance against negative feedback (antisocial behavior). The experiment outcome also supports this hypothesis. Another factor we explore is an ingroup framing of the task. People show a higher level of empathy towards ingroup members [38]. In our work, we pair participants and offer them a group-based reward. Prior work finds a temporary framing stimulates ingroup oriented behaviors [39]. In our study, we explore whether such a framing stimulates a sense of in-group membership between the feedback provider and the designer, which arouses empathy and leads to more prosocial behaviors. As we report in the latter part of this thesis, the data analysis confirms in-group framing could indeed increase feedback quality.

1.3 VISION AND A USE CASE SCENARIO

In an ideal world, content creators can always rely on motivated feedback providers for constructive yet considerate advice. But, this is not always the case in reality. It is not uncommon to see aggressive feedback on online platforms. Currently, content creators are often left on their own to develop strategies to cope with these negative feedback. This approach is undesirable for a few reasons. First, content creators experience emotional distress in the process of developing these skills. Also, it may take years to become resilient to negative feedback; and many content creators may never be able to cultivate enough resilience. Novice content creators who have not faced negative feedback before may also be discouraged from pursuing their interests in design. Unfortunately, novice content creators are often more prone to negative feedback due to their lack of expertise and relatively low quality of the shared work. Because of these shortcomings of the current approach, I proposed a series of mechanisms that facilitate content creators to cope with negative feedback and

reduce the likelihood of receiving such feedback at the same time. Below I illustrate how my interventions achieve this goal in a use case scenario.

Imagine a novice and inexperienced content creator, who has just invested significant effort into a graphic design that she feels very confident about, visits an online design community seeking feedback. After she posted the design, a few community members find the post and review the work. To affirm the members' commitment to providing constructive feedback, the platform labels the feedback provider and content creator as a team to shift feedback providers' mindset into a more collaborative one. Feedback providers also read a short narrative of the design process of the work, which further aligns their perspective with one of the content creators. After knowing more about the content creator and now regarding her as a teammate, the community members became notably more motivated to help the content creator by providing high quality feedback.

While most feedback is constructive, an uncooperative user ignores community guidelines and proceeds to question the skills and motives of the content creator . The platform quickly detects the negative feedback using off-the-shelf sentimental analysis algorithms. To dampen the impact of the negative feedback, it changes how the feedback is presented and reveals the negative feedback at the end, after presenting the pieces of feedback with more positive sentiment. The affirmation from the positive feedback acts as a buffer against the negative feedback. As a result, while the content creator still feels slightly distressed, reading the constructive feedback significantly reduce the adverse effects of the negative.

As the content creator is still moderately distressed, the platform prompts her, in a post after all collected feedback that is only visible to her, to revisit the earlier feedback and then write down her thoughts and feelings at the moment. This short period of reflection helps her to gain a deeper understanding of the feedback and leads to more improvement in the revision.

This scenario envisions how all the techniques might integrated into a single feedback collection experience. However, in practice, we would expect that platforms would implement different subsets of these techniques.

1.4 CONTRIBUTIONS

My thesis proposes a set of solutions that control negative feedback from two new directions: mitigating the influences of negative feedback on content creators and increasing feedback providers' intrinsic motivations. Specifically, my work makes the following contributions:

1. Examine three theory-based coping activities and provide empirical evidence on how they mitigate the effects of negative feedback over different measures: Our results show even a small amount of negativity has significant adverse effects on all the measures. For the coping activities, we find that expressive writing encourages essay revision, distraction improves affective states and feedback provider perception, and self-affirmation has no significant effects on the measures. Our results contribute further empirical knowledge of how negative valence feedback impacts content creators and how the coping activities tested mitigate these effects. We also offer practical guidelines regarding when and how to use the activities tested in online feedback platforms. These findings were published in CSCW 2018 [40].

2. Create a novel valence-based feedback ordering mechanism that mitigated the effects of negative feedback: Our main result is that presenting negative feedback last improves content creators' affective states and their perception of the feedback set relative to placing the negative feedback in other positions. This pattern stays consistent across all feedback source conditions, including experts, peers, and anonymous source. The work contributes a simple and novel way to order a set of feedback that improves feedback receptivity. These findings were published in C&C 2017 [41].

3. Explore whether empathy arousal and ingroup framing can encourage feedback providers to compose more friendly and higher quality feedback: our results showed that both narrative based empathy arousal methods and ingroup framing interventions significantly increased the effort invested in feedback composition by 40% and the final feedback quality by 30%. We also examined an empathy arousal method where the content creator shares her past experience receiving negative feedback. Participants who read this narrative report a significantly more disapproving stance towards negative feedback and are significantly more likely to intervene when negative feedback occurred. These findings have been accepted at CSCW 2020 [42].

4. 4. Provide empirical evidence that negative information about the feedback provider significantly lowers people's perception of the feedback and its providers: We investigate two information cues about feedback providers – the effort invested in a feedback task and expertise in the domain. First, we test how positive and negative cues of a provider's effort and expertise affect the perceived quality of the feedback. Results show both cues affect perceived quality, but primarily when the cues are negative. The results also show that effort cues affect perceived quality as much as expertise. In a second study, we explore the use of behavioral data for modeling effort for feedback

tasks. For binary classification, the model achieves up to 92% accuracy relative to human raters. This result validates the feasibility of implementing effort cues in crowd services. These findings were published in CHI 2016 [43].

Overall, my thesis explores two new categories of interventions that address the problem of negative feedback: one category of interventions aimed at mitigating the impacts of negative feedback over content creators. And the other category aimed at increasing feedback providers' intrinsic motivations for writing constructive feedback. We believe these techniques will progress toward a future where content creators at all skill levels are able to improve their creative projects and skills by acquiring constructive feedback online and being more resilient when the feedback is not.

CHAPTER 2: RELATED WORK

In this chapter, we discuss related work for different mechanisms included in my thesis. Section 2.1 discusses how we define negative feedback and what role feedback valence plays in creative work; Section 2.2 describes existing approaches to control or mitigate negative feedback; Section 2.3 explores state-of-the-art sentimental analysis algorithms that would enable the mechanisms I propose in this thesis; Section 2.4 elaborates on the theoretical bases of the coping activities in Chapter 4; Section 2.5 covers prior work regarding the use of positive valence feedback and ways to convert feedback into more positive ones, which relates to the feedback reordering mechanism in Chapter 5; Section 2.6 explains how empathy arousal and ingroup framing may improve the user experience of feedback exchange, providing a foundation for our approach in Chapter 6; lastly, section 2.7 covers the related work for the preliminary study discussed in Chapter 3, explaining how information cues influence users' perception of feedback.

2.1 INFLUENCE OF NEGATIVE FEEDBACK

Content creators sometimes receive feedback delivered in an overly negative tone, which we refer to as negative feedback in this dissertation. A piece of negative feedback can target either the content creator himself, such as the example in Figure 1.1, where the feedback is unnecessarily harsh in relation to the perceived experience of the content creator. The negative feedback can also target the content of the design. For example, on a popular feedback exchange platform, a novice content creator seeking constructive feedback was told the work is "terrible for a ton of different reasons" without receiving specific advice on improving . This negative feedback significantly reduces the content creators' affective states and the quality of the design solution [40]. This issue is not uncommon, as users with antisocial tendencies disproportionately generate negative valence content [44, 45]. Furthermore, negative feedback snowballs [46]. One piece of negative feedback can incite more people to provide similar feedback. Due to the above-mentioned issues, feedback exchange platforms frequently fail to deliver constructive feedback.

Researchers have explored various factors that influence feedback receptivity, including granularity, modality, and timing [19, 47, 48]. For example, Sadler argues effective feedback needs to address the discrepancy between the current performance level and the desired goal [49]. Valence is another factor with a strong influence over content creators' reactions to the feedback received. Negative valence language threatens one's ego and reduces feedback ef-

fectiveness [50]. Positive language typically improves feedback reception, task performance, and content creators' affective states [27]. Zhu et al. show negative feedback discourages participation in online production communities while positive feedback encourages participation [51, 52].

2.2 EXISTING APPROACHES TO ADDRESS NEGATIVE FEEDBACK

Before discussing related work for each part of this thesis, we would like to describe how prior work addresses the issue of negative feedback from various perspectives.

One perspective is to curb negative valence content by controlling its source, i.e., who has the permission to provide feedback. In practice, many online platforms use blacklist/whitelist maintained by users and moderators to control the feedback source. Prior work shows a small group of online users with antisocial tendency generate more negative valence content than users without such tendencies [44, 45]. To limit the influence of these users, platform designers may attempt to block them from future participation [53]. Many online communities rely on either centralized [54] or distributed moderation [9, 10] to identify users with such characteristics. While centralized moderation holds a more consistent standard, distributed moderation is easier to scale up. For both approaches, prior work shows such moderation facilitates community interaction and helps to sustain growth [54]. Platform designers can also implement reputation systems to assist community moderation. Users are less likely to exhibit antisocial behaviors when the system logs malicious actions [55]. A reputation system also makes it easier for moderators to identify users who habitually generate negative content. However, a recent study shows users with a benign feedback history may also contribute negative feedback when experiencing negative moods or after exposure to negative valence content [5]. Existing practices targeting users with antisocial behaviors are less effective in these situations.

Another way to control negative feedback is by controlling how the feedback is composed. Although content creators have limited control over this aspect, they can still set a bound over how negative feedback can be by specifying the desired feedback type [15]. On the feedback provider end, prior work uses rubrics [13] and model feedback [14] to encourage them to provide high-quality feedback. One drawback of these interventions is their dependence on users' intrinsic motivations to provide good feedback and may be ineffective if users choose to ignore the interventions. Platform designers can also build a scaffolding process to strictly limit how feedback providers compose the feedback in order to reduce the potential negativity of the content [15, 17]. However, such methods also sacrifice the potential depth and usefulness of feedback at the same time.

A third way to address the problem of negative feedback is to control who receives the feedback. Online platforms can rely on central/distributed moderation [9, 10] and learning-based filtering [21] to identify and block negative feedback. Alternatively, platforms can also use tags to label content inappropriate for specific user groups, including under-age group, inexperienced users within the community, etc. While such methods are common for controlling negative content, so far, no prior research has used them to control negative feedback.

Existing approaches also try to control negative feedback by specifying how the feedback is presented. Prior work uses machine learning models to empower content creators to select the type of feedback they want to review [23, 24]. Another method to control the impact of negative feedback is via feedback aggregation, which creates a buffer between the negativity in the feedback and the content creator [26]. While this approach succeeds in preventing negative feedback from reaching the content generators, it only works when there is a large volume of feedback available. Alternatively, platform designers and peer feedback providers can also help to censor negative feedback before content creators read it. Such moderation can be achieved either by distributed moderation [9, 10] or machine learning-based classification methods [21].

At last, platforms can also use mechanisms to control how the negative content will be consumed. Existing approaches tag specific topics, language style, and phrasing. Afterward, content receivers could decide whether they want to consume certain tagged content or not.

2.3 REVIEW ON SENTIMENT ANALYSIS TECHNIQUES

The coping activities and the feedback reordering mechanisms proposed in my thesis rely on the support of off-the-shelf sentimental analysis algorithms. Platforms need to detect negative feedback first so deploying mitigation systems will not impede normal feedback exchange. Identifying one piece of negative feedback among all received ones enables valence-based feedback order. Monitoring feedback negativity trends also help to indicate how healthy community interactions are and gauge the need for mitigation methods. In this section, we report the performance of the state-of-the-art sentimental analysis algorithms in order to demonstrate the feasibility of the proposed methods.

While traditional approaches have acquired satisfactory results on sentimental analysis, recent advancement in deep learning has further increased the accuracy to a near-human level. Traditional methods, including TFIDF based on bag-of-words [56], TFIDF based on bag-of-n-grams, and word embedding [57], have reached 80∼90% accuracy as reported in Zhang et al.'s work when identifying positive (rating 4 & 5 on a five-point scale) and

negative (rating 1 & 2) Amazon product reviews [58]. Furthermore, recent deep learning-based work has reached ∼98% accuracy for binary classification (positive vs. negative) of restaurant reviews collected from Yelp.com [59, 60, 61, 62, 63]. While restaurant reviews may be more objective than a critique on creative work and thus easier to identify the dominant sentiments, recent work has also reached a satisfactory level of accuracy (∼96%) for movie reviews, a form of design critique, with intense sentiments (ratings significantly higher or lower than the average) [59, 64, 65, 66]. The results discussed here help to show the off-the-shelf algorithms are effective enough for the proposed methods in our work.

While existing work has achieved satisfactory results with binary classification, sentimental analysis at a higher granularity (five-class vs. binary) is much more challenging. State-of-the-art algorithms could only achieve ∼70% accuracy for fine-grained five-class classification on restaurant reviews [59, 60, 61, 62, 63]. For movie reviews, the task turns out to be more challenging as the state-of-the-art algorithms could only achieve a ∼55% accuracy. Both the coping activities and the feedback reordering mechanisms rely only on a binary classification to function properly. However, a more accurate fine-grained classification may open doors to other intervention mechanisms. Platform designers could rely on a fine-grained classification to fine-tune the threshold for initiating coping activities, since the interventions may require more effort from the feedback receivers. In addition, for feedback reordering, fine-grained classification allows us to further examine the effect size of affirmation provided by the feedback of different levels of positivity.

Another relevant problem in sentimental analysis is aspect-based analysis [67]. One piece of feedback may have conflicting views on the same design, approving decisions on one aspect while disagreeing with the ones on another. State-of-the-art algorithm has achieved decent accuracy regarding aspect identification (F1 score 70∼80) and polarity detection (∼90%) [60, 68, 69]. Being able to identify whether feedback from different sources converge upon the same aspects allow us to gain more insights into the effectiveness of manipulations. It remains as an open question for future study to explore, that whether affirmative feedback needs to counter the negative feedback on the same aspects to be effective, or whether convergence on the same aspects may lead to larger effect sizes.

2.4 THEORETICAL FOUNDATIONS OF THE COPING ACTIVITIES

In this section, we discuss the theoretical foundations for the coping activities in Chapter 4. We also explain why we think these activities could help address the problem of negative feedback.

2.4.1 Self-Affirmation

Prior work shows self-affirmation is an effective ego-protection mechanism [31, 70]. Facing information that threatens self-integrity, people are more likely to react defensively and become less receptive [31]. Affirming people's self-worth before exposure to ego-threatening information deters them from taking defensive measures [30]. The affirmation on people's core values serves as a buffer against the information that threatens the perceived integrity of the self. Prior work shows self-affirmation reduces resistance to disconfirming evidence in social-political discussions [71], negative valence health-risk information [72, 73], and critical feedback on a public speaking task [74]. In online feedback collection, the affirmation of core values may neutralize the effects of negative feedback. In Chapter 4, we hypothesize self-affirmation can help participants to preserve their self-worth in the face of negative feedback and maintain positive affective states. At the same time, self-affirmed participants may stay receptive to the constructive critiques despite the negative tone of the feedback.

2.4.2 Expressive Writing

In expressive writing, people recognize their current emotional states and express them in written form [75]. The activity allows people to cope with stress by reexamining and reinterpreting their experiences via writing [76]. Prior work reports expressive writing reduces students' anxiety level during exams for high test anxious students [77]. On the other hand, writing helps students to reflect on the feedback they received and increases task performance [78]. Feedback evokes content creators to reflect on their design, and the writing process facilitates a deeper level of contemplation. In our work, we hypothesize that expressive writing can help participants recognize and process their emotions and lead to more positive affective states after they read the negative feedback. Meanwhile, the writing process may also stimulate participants to reexamine the feedback and gain insight.

2.4.3 Distraction Intervention

Distraction can attenuate depressive moods [79] and relieve anxiety [80]. It is a common coping strategy that people frequently initiate to abstain from brooding over existing problems. Focusing on neutral or pleasant tasks occupies people's cognitive load and stops them from rumination. Distraction may help content creators recover from emotional discomfort and increase creativity [81]. In addition, a short duration of mind wandering facilitates creative problem solving [36]. Thoughts generated during the distraction may help people

to view the existing problem from a new angle. In an online feedback collection process, performing an unrelated task could stop people from ruminating over the negative feedback and improve their affective states. The distraction may also stimulate creative thoughts and lead to higher quality revision.

2.5 THE IMPACTS OF POSITIVE VALENCE FEEDBACK AND FEEDBACK SOURCE IDENTITIES

In this section, we discuss prior work regarding how positive valence content and source identities affect feedback consumption. These prior studies serve as a foundation for the proposed feedback reordering mechanism in Chapter 5.

2.5.1 Using Positive Valence Feedback as a Buffer

Feedback valence is particularly important for creative tasks. Positive affective state, which can be affected by feedback, relates to improved creativity [81, 82, 83]. Researchers have explored workflows that utilize the effects of positive valence in creative work. Nguyen et al. modify the valence level of the feedback by inserting positive affective language at the beginning of the text [27]. Such a positive language "wrapper" was shown to improve feedback reception and writing quality. De Rooij & Jones show that displaying positive feedback in real-time can encourage participants to generate more original ideas [84]. Prior work has also explored how manipulating the valence of phrases within a single piece of feedback influences content creators. One prior study shows that delivering negative feedback with positive feedback framing increases its perceived usefulness and participants' confidence but does not affect participants' performance on a repetitive physical task [85]. Prior work has also shown that mitigating language increases participants' receptivity to feedback and their affective state [86, 87]. In contrast, our work explores the effects of valence ordering in the context of multiple pieces of feedback and in the context of a creative writing task.

2.5.2 Source Identity's Influences on Feedback Receptivity

Besides the valence of the feedback, we also study the identity of feedback providers in Chapter 5, which is another common factor platform designers use to categorize feedback. Online work marketplaces, such as UpWork, allow content creators to identify and collect feedback from paid domain experts. Feedback from experts leads to greater improvement in technical skills than self-assessment [88]. Prior work also finds content creators are more

likely to accept feedback from experts because of their high perceived credibility [89]. On the other hand, collecting feedback from peers is also gaining popularity among content creators [90]. Although less experienced than experts, peers are typically more accessible for feedback collection. Prior work also shows peers are more responsive and provide more design suggestions in comparison with online design forums [1].

Besides expert and peer sources of feedback, in our work, we also explore an anonymous condition, where no identity information about the source is given. Anonymity removes the social interaction element in the feedback interpretation process. Prior work shows the lack of social cues increases participants' motivation, perceived ability, and task performance when receiving computer-generated feedback [91]. Nguyen et al. show anonymity also increases feedback acceptance [27]. Our study extends this corpus of prior research by testing how feedback valence order interacts with different source identity cues.

2.6 EMPATHY AROUSAL METHODS AND THEIR INTERACTIONS WITH INGROUP FRAMING

In this section, I discuss prior work on empathy arousal and how it affect prosocial behaviors. I also survey literature regarding in-group framing and how it mediates empathy arousal in prior empirical studies. These works provide a foundation for the interventions explored in Chapter 6.

2.6.1 Empathy-Inducing Narratives and Their Influences on Helping Behaviors

After exploring methods that mitigate the impacts of negative feedback in Chapter 4 & 5, including the coping activities and feedback reordering, the results indicate that it is challenging to completely negate the influences of negative feedback. Thus, in Chapter 6, I explore a new technique based on the theory of empathy arousal to encourage users to write feedback with a more constructive tone. As we discussed earlier, prior research has explored methods to reach this goal, but these methods sometimes fall short because of a lack of intrinsic motivations from the feedback providers. Moreover, prior work shows much of the aggressive content online is provided by ordinary users in negative affective states [5]. For this type of users, empathy arousal may help them to develop the motivation to contribute constructive content rather than venting. Therefore, we examine whether empathy arousal may increase feedback providers' intrinsic motivation.

Empathy is a vicarious response to others' emotional states [92]. Prior work has explored various empathy arousal methods [93, 94, 95, 96, 97, 98]. One of the most common and

more online-appropriate ways to elicit empathy is via narratives. Prior work shows habitual fiction readers report a higher than average level of empathy [99]. Reading narratives also has immediate effects on people's empathy. Prior work shows reading a short narrative essay can lead to higher empathy and higher prosocial behavior immediately after performing the task [99, 100]. Empathy towards people in distress causes an emotional appraisal and leads to prosocial behaviors [101]. While researchers haven't reached a consensus regarding the definition of empathy, most agree that it includes both a cognitive part and an affective part [102, 103]. The former operates when people analyze the target's experience and current situation to deduce his emotions at the moment [102]; the later operates when people intuitively recognize the target's emotions [102]. In our experiment, we plan to test whether two types of narratives corresponding to these two aspects of empathy: negative experience and design process.

For the negative experience narrative, participants review the design accompanied by a narrative about the designer's recollection of receiving negative feedback. Prior work shows sharing unpleasant experience induces empathy and encourages helping behaviors [104]. Also, reading narratives about unpleasant experiences discourages people from inflicting similar experiences to others and make them less tolerant about the offensive behaviors [105]. In our study, the negative experience narrative should encourage participants to perform more helping behaviors, which is to provide more useful feedback in the context of feedback collection, and increase their tendency to intervene when they observe offensive behaviors from other users, such as providing negative feedback. In the design process condition, the narrative includes a description of the goal of the design and a series of explained design decisions made in the process. Prior work shows information about the protagonist of the narrative, including their background, goal, and their journey so far, cultivates empathetic feelings in the readers [106, 107, 108]. In our study, we try to frame the design process in a similar way to arouse content creators' empathy. The narrative describes the goal of the design, along with how the designer planned to achieve it and his current progress. In addition, the design process narrative also has the potential to increase the usefulness of the feedback. Prior feedback theory argues a piece of feedback needs to accurately identify the goal of the design, the current state of the design, and actionable advice to reach the goal, to be constructive to the content creators [49]. The design process narrative helps the content creators to judge these aspects of the design more accurately.

2.6.2 Interaction between Ingroup Framing and Empathy Arousal

Another intervention we examine in Chapter 6 is ingroup framing. Perceiving others as ingroup members makes it easier for people to develop empathetic feelings and show proso-

cial behaviors. Prior work shows people feel more empathetic and have stronger prosocial tendencies towards ingroup members [109]. Empathy towards ingroup members may even induce costly helping behaviors, such as choosing to endure physical pains for other people [110]. Prior empirical studies also suggest people may show empathy only towards ingroup members [38, 111]. In extreme cases, people may not only fail to feel empathy towards outgroup members but instead gain pleasure from the suffering of outgroup members who they dislike [112]. Fortunately, such differences in attitudes are not immutable. Prior work shows changes in social categorization influence group membership perception and the likelihood of empathy arousal [113]. Researchers had also shown empathy arousal interventions could be used to improve intergroup relationships [114]. In our study, we are less interested in the difference between ingroup and outgroup members and would like to focus on using ingroup framing to arouse empathy and increase the likelihood of prosocial behaviors. Prior work argues group labeling and interdependent relationships foster a sense of group membership [115]. Prior empirical study also demonstrates such interventions may stimulate participants to view computer agents as their teammates [39]. In our study, we would like to test similar interventions and their influences in feedback generation tasks.

2.7 HOW INFORMATION CUES INFLUENCE FEEDBACK PERCEPTION

In this section, I discuss prior work on factors that affect content creators' evaluation of the feedback received. This work serves as a foundation for the preliminary study.

2.7.1 Assessing Online Content

Relevant cues provided in the information environment can help users better judge credibility [116, 117], weigh conflicting views [118], make decisions [119], and prioritize suggestions. An important and generalizable cue is the expertise of the content's author. Researchers have studied how this cue relates to content assessment [116, 117, 118]. For example, Liao and Fu found that online comments showing indicators of high expertise were selected by users for reading more often than comments without such indicators [118]. In a large-scale study, Fogg et al. found that the presence of expertise cues related to more favorable perceptions of website credibility [116]. In our work, we are also interested in how expertise cues affect the assessment of online content. However, our work tests the effects of expertise cues for evaluating online design feedback, a unique type of content; how cues of the provider's expertise interact with cues of his or her effort for the assessments; and how these effects are mediated by the intrinsic quality of the feedback.

Researchers have also identified the need to consider the effort of the content's author when assessing its quality, especially for crowdsourced work [24]. For example, one crowdsourcing study reported that up to 50% of the responses were of poor quality due in part to workers not investing sufficient effort into the task [120]. Our work is the first to study how explicit cues of effort affect the assessment of the quality of crowd work. Many other cues have also been studied for assessing content online [17, 117, 118, 121], but our focus is on studying the cues of effort and expertise in a crowdsourcing context.

2.7.2 Social Transparency in Online Work

Social transparency is defined as the availability of social meta-data surrounding information exchange [122]. Receiving design feedback is one form of information exchange and therefore it can be situated in the framework of social transparency. Though social transparency points to many attributes of social meta-data, our work considers two: expertise and effort. Expertise can be regarded as an attribute of identity transparency because it reflects a person's knowledge in the domain of interest. Effort can be regarded as an attribute of content transparency because it relates to the provider's behavior around the creation of the feedback. We prioritized these two attributes because expertise has been shown to be important for assessing online content [118] while effort has been described as being critical for interpreting crowdsourced work [24].

Prior work indicates that increasing social transparency can improve the quality of crowdsourced work [123] and affect impressions of those who performed the work [17]. However, the focus of these prior studies was to increase the transparency between crowd workers, whereas we are increasing the transparency between a designer (requester) and the feedback providers (workers) to help the designer better interpret their responses. We are also using different transparency cues that are relevant to a design context.

2.7.3 Modeling Crowdsourced Work

There is growing interest in modeling the behavior of crowd workers for improved quality control [23, 24], task pricing [124], and activity history [121], among others. For instance, Rzeszotarski and Kittur have shown that behavioral traces of workers can be leveraged to predict response quality [24]. The authors further showed that models of behavior could be used to cluster workers who share similar patterns of work [23]. In contrast, part of our work tests how well models of behavior can be used to predict perceived effort rather than response quality. To determine a fair price for crowd work, Cheng and Bernstein leverage

the objective performance data of workers to measure the intrinsic difficulty of a task and use it to set the task's price [124]. We are using the way workers perform a feedback task, which is subjective and open-ended, to model the perceived effort invested by the worker rather than the intrinsic difficulty of the task with the goal of helping designers better assess the feedback. Researchers have also developed models of crowdsourced work for completing work under budget or time constraints [125] or recommending tasks [126]. In contrast, our focus is on using behavioral traces to model perceived effort and to study its impact on judgments of the quality of crowd work.

CHAPTER 3: PRELIMINARY STUDY

Crowd feedback services offer a new method for acquiring formative feedback during the iterative design process [127]. The services utilize online crowds as a simulated audience to collect, aggregate, and present their interpretation of design [14, 127, 128]. Relative to soliciting feedback from peers and online communities, the benefits of these services include the ability to acquire feedback on-demand without burning social capital or needing online reputation [121], the integration of scaffolding to boost feedback quality [128], and access to a diverse and scalable audience [127]. Crowd feedback services can be used to acquire feedback on Web, product, and interaction designs, among other genres.

An empirical study of one representative service found that crowd feedback helps designers improve their designs in an iterative process [15]. However, in that study and other work [128], designers reported wanting to know more about the providers giving the feedback. This information could be used for assessing the credibility of responses, weighing conflicting viewpoints, and prioritizing suggestions. The problem is that existing crowd feedback services only show the feedback, without any information about the providers. One key reason is that there is little empirical knowledge about what information these services should display.

In this section, we draw on social transparency theory to study how presenting two critical cues about the providers – their effort and expertise – affect the perceived quality of their feedback. In this chapter, effort is how much energy a provider invests in performing a crowdsourced task. For example, for a design feedback task, effort may include how long a provider views the design, length and number of revisions of the text, and the precision of the annotation. Expertise refers to the level of domain knowledge. While expertise has been studied for assessing online content [117, 118] and effort has been cited as critical for assessing crowd work [24], our work synthesizes and investigates these two cues for interpreting crowdsourced design feedback.

Our investigation of these cues consisted of two studies. In the first study, we generated an authentic dataset of design feedback and asked human raters (N=2700) to review each response and rate its perceived quality. In the rating interface, we manipulated a block of text giving positive and negative cues of the effort and expertise of the feedback provider. Results showed that both cues affect judgments of perceived quality relative to a baseline condition (up to 21% difference), but mainly for the negative manipulations. Surprisingly, we also found that indicating effort affects the perceived quality ratings as much as indicating expertise.

The results argue for implementing these cues in crowd feedback services, e.g., to help designers interpret and differentiate the feedback. For expertise, system designers can choose between several existing methods (e.g., [24, 129, 130]). However, implementing cues of effort is challenging because it is a task-specific behavior, and there has been little research aimed at measuring it for crowdsourced tasks.

Our second study therefore addressed this gap. We first collected behavioral traces of providers performing three feedback tasks. Through a software tool that we developed, human raters viewed replays of the workers performing the tasks and rated the perceived effort. A novel aspect of our replay tool is that it masked the characters during text entry to focus attention on the behavior rather than the content. Statistical models were built to learn mappings from the behavioral data to the perceived effort ratings. For a binary classification of effort, the models achieved 92% accuracy. This outcome demonstrates the feasibility of implementing effort cues within feedback services and other crowd tasks.

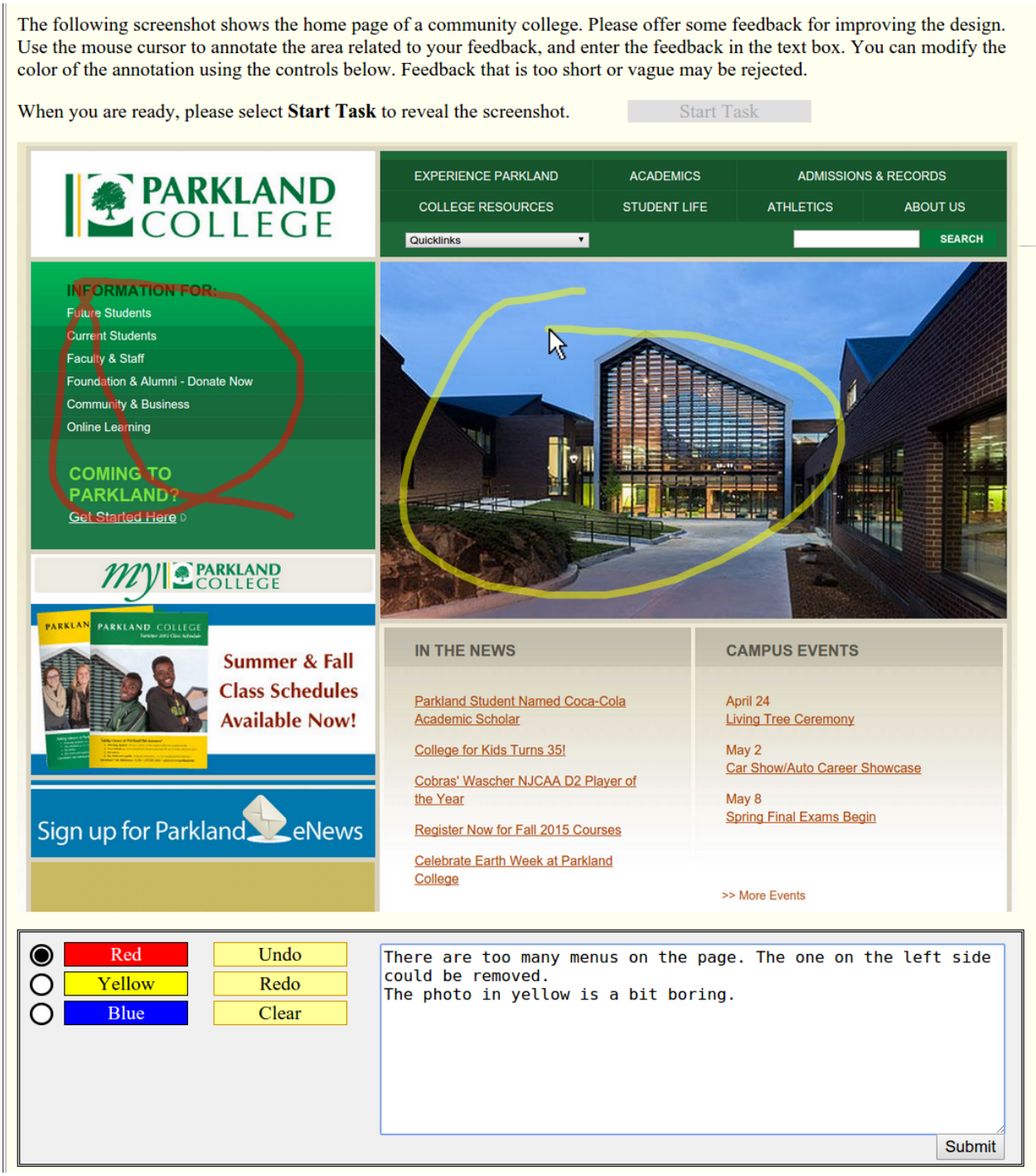
3.1 STUDY ONE: METHODOLOGY

We first investigate how different information cues about the feedback providers influences the reception of the feedback. Throughout the feedback collection process, content generators receive various information, some positive and some negative, about the feedback providers. In this preliminary study, we would like to examine how the two most common information cues, i.e., effort and expertise, affects content generators' perception of the feedback. The study addresses two fundamental research questions:

- RQ1: How do explicit cues of a provider's effort and expertise affect the perceived quality of the feedback provided for a design?
- RQ2: How are the effects of these cues mediated by the intrinsic quality of the feedback?

There is a large space of potential cues (social activity, demographics, geography, etc.), but we prioritized two to keep the study tractable. Expertise was included because it has been shown to be important for assessing online content [118] and because designers have reported wanting this cue for interpreting crowd feedback [15]. Effort was included because it has been previously described as important for assessing the quality of crowdsourced responses [24].

Figure 3.1: The user interface for collecting authentic design feedback from the crowd. Providers reviewed the Web design and left feedback using the edit box and annotation tools.



3.1.1 Experimental Design

To answer these questions, we conduct a full-factorial, between-subjects experiment. The factors are Effort (High vs. Low vs. Not Given) x Expertise (High vs. Low vs. Not Given) x Intrinsic Quality (low=1 to high=5), giving a 3x3x5 design. The experimental design and manipulations draw from similar studies testing how informational cues affect judgments in other domains (e.g. [55, 82, 106, 131]).

3.1.2 Design Feedback Dataset

For the experiment, we generated an authentic dataset of design feedback from feedback providers and developed an intrinsic quality score for each response. The design was the home page of a community college (http://parkland.edu). It was selected because its content should be familiar to a general audience, it was not too complex, and there were many opportunities for design improvements.

Feedback providers were recruited from Amazon.com's Mechanical Turk. A HIT was posted asking the providers (workers) to inspect the page and describe how it could be improved. The instructions also stated that feedback that was too short or vague would be rejected. As shown in Figure 3.1, the feedback collection interface included a screen capture of the page, an edit box for entering text, and a free-form ink tool (added via JavaScript) for annotating image regions corresponding to the text. The ink tool supported multiple colors and operations such as undo and clear. The interface was designed to simulate existing crowd feedback services. Sixty pieces of feedback were collected, each from a different provider. A provider received $0.50 (US) and was required to have a 95% prior approval rating.

Three judges with experience in HCI were recruited from our institution to review the design and then rate the quality of each piece of feedback. The judges had no affiliation with this research project. The rating was performed on a 5 point Likert scale from 1 (lowest quality) to 5 (highest). For calibration, each judge first reviewed a sample of the feedback at different quality levels based on our own analysis and was encouraged to use the full range of the scale. A judge viewed the feedback online, one response at a time, and entered ratings in an online spreadsheet shown on a second monitor. A judge could review the feedback and modify the ratings until satisfied. Each judge completed the ratings in about one hour and received $15 (US).

Once the ratings were collected, we averaged the three ratings for each feedback response and rounded to give the final classification, or intrinsic quality score. On the scale of 1 to 5, the distribution of the classifications was 16, 14, 16, 12, and 2 respectively. Krippendorff's

alpha, a measure of reliability for multiple raters and categories, was 0.71, which represents good agreement [132]. The feedback with higher intrinsic quality scores typically had more words (μ=633.0 for level 5 vs. μ=77.9 for level 1), suggested more improvements, and the suggestions were more specific and actionable. One feedback response from each level of intrinsic quality was randomly selected for the experiment. The feedback selected for the study is shown in Figure 3.2.

3.1.3 Experiment Interface

The rating interface showed a feedback response, a block of text about the feedback provider, and interaction for rating the perceived quality of the feedback. Perceived quality was rated on a scale from 1 (low) to 5 (high). See Figure 3.3.

The cues for effort and expertise are manipulated in the block of text, and follow a similar linguistic pattern. For expertise, the pattern is: "It is known that the person who left the feedback has $LEVEL knowledge of design." Effort follows a similar pattern: "It is known that the person who left the feedback invested $LEVEL effort to develop the feedback". $LEVEL is replaced with "minimal" and "significant" in the respective conditions. For example, if Low Effort is crossed with High Expertise, the block of text will read: "It is known that the person who left the feedback invested minimal effort to develop the feedback. It is also known that the person who left the feedback has significant knowledge of design." For a Not Given condition, the respective sentence is not included. The block of text is a manipulation in the experiment and is not related to the provider who actually leaves the feedback. The blocks of text for each level of effort and expertise are then replicated for the feedback representing the five levels of intrinsic quality. All 45 conditions are constructed a priori. When neither sentence is provided (i.e. effort not given and expertise not given), the participant only sees the feedback and the instructions for rating it, thereby serving as the baseline condition for the feedback at each level of intrinsic quality.

3.1.4 Participants

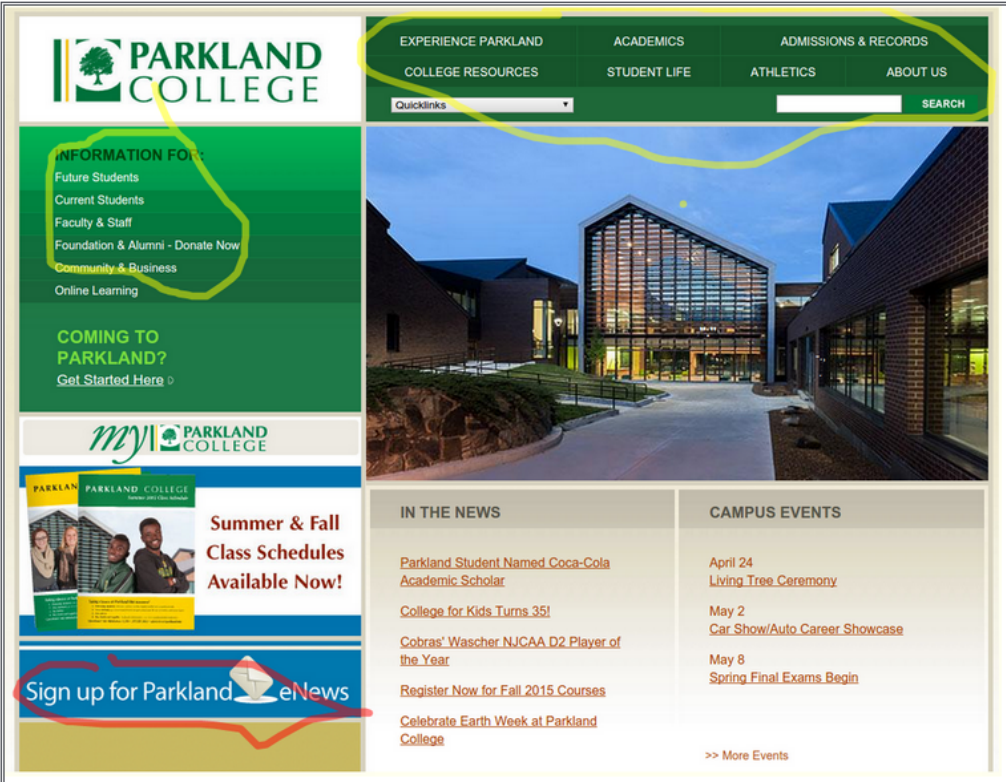
Participants (N=2700) are recruited from Mechanical Turk. Participants reside in the US (84.1%), India (12.1%), and 46 other countries (3.8%). We do not anticipate age or gender effects, and therefore do not collect this demographic data to minimize privacy concerns and to reduce the overall length of the task (HIT).

Figure 3.2: The authentic design feedback sampled at each level of intrinsic quality for Study 1. Some text in the top row (IQ 5) was omitted for brevity. Participants rated the perceived quality of the feedback with manipulations of the effort and expertise of the provider who supposedly left each response.

IQ	Feedback Text	Annotation
5 (High)	The "current students" section could easily be placed under the "student life" section. That's where I always found it for the homage of my university. I would also place "online learning" under academics … What is the difference between the two blue areas I circled? … Register Now should be at the very top of the page. …	
4	Red - I don't see the relation in the quick links and having other "quick" links at the top. Yellow - Font could be larger to be more appealing. Blue - "Important" information seem bland; could be presented a little more interactive to increase circulation.	
3	Make the information bar and the top bars more visible. They are boring and need to be more interesting to the student. Also get rid of the sign up for parkland enews since that should put on a page for current students. make your page more demanding for the prospective students.	
2	In ability to add campus events to an existing calendar, such as google calendar or iPhone.	
1 (low)	Styling is not good.	

Figure 3.3: The interface for rating the perceived quality of the design feedback. It shows a feedback response, a block of text (manipulated) about the provider, and the rating interaction.

Please review the feedback provided for the community college page below. The feedback consists of annotations on the image and comments below the image.



Make the information bar and the top bars more visible. They are boring and need to be more interesting to the student. Also get rid of the sign up for parkland enews since that should put on a page for current students. make your page more demanding for the prospective students.

Task: It is known that the person who left the feedback invested significant effort to develop the feedback. It is also known that the person who left the feedback has little to no knowledge of design. From 1 (least useful) to 5 (most useful), rate the perceived usefulness of the feedback.

1	2	3	4	5

3.1.5 Procedure

Upon accepting the task, the participant is randomly assigned to one of the 45 conditions. Each condition is shown using the experiment interface previously described. The participant is instructed to review the feedback response for the design and rate its perceived quality (1 to 5). The manipulation of the text about the provider is integrated into the rating request, which is displayed below the design feedback and general task instructions. Participants therefore read the manipulation about the provider after viewing the feedback, but before rating it. Based on pilot testing, this placement achieves high likelihood that the manipulation is read. A participant receives $0.35 (US) for performing the task. It is configured to require 95% prior approval and to allow workers to only participate once. The batch is posted on AMT from June 29th to July 13th, 2015.

3.2 STUDY ONE: RESULTS

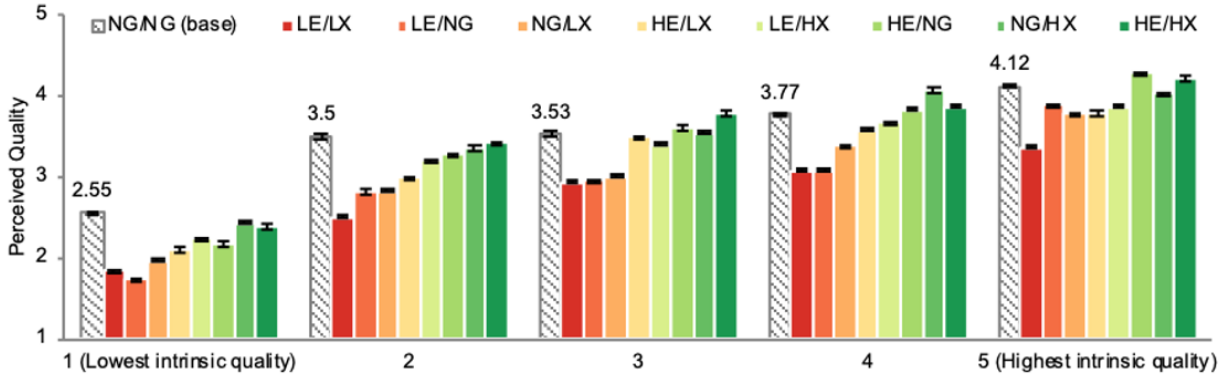
We collect 3,081 ratings and discard 381 redundant or corrupted ones due to technical issues. This leaves 2700 ratings for analysis, 60 per condition. The perceived quality ratings are shown in Figure 3.4 clustered by each level of intrinsic quality. The ratings are analyzed using a 3-way ANOVA with Effort, Expertise, and Intrinsic quality as factors. Bonferroni corrections are applied to pairwise comparisons to control for family-wise error. The statistical results are summarized in Table 3.1.

Results show an interaction effect between effort and expertise. Relative to the mean of the five baseline conditions (μ=3.49), signaling low effort regardless of expertise lowers the mean rating of perceived quality (μ=2.97, $p<0.001$). Signaling low expertise regardless of effort also lowers the mean rating (μ=2.97, $p<0.001$). However, when these cues are combined, the ratings of perceived quality are the lowest (μ=2.74, $p<0.001$). In this case, the perceived quality of the feedback is reduced by 21% relative to the mean of the baselines. This pattern is consistent for the feedback at each level of intrinsic quality (see Figure 3.5) and is further supported by the lack of a three-way interaction.

Interestingly, the mean of the Low Effort / No Expertise Given condition (μ=2.89) is lower than the No Effort Given / Low Expertise condition (μ=2.99, $p<0.001$). Knowing that a feedback provider does not invest effort into the task reduces perceptions of the work quality more than knowing that the provider only has minimal knowledge of the domain.

It is also surprising that when positive cues about a provider were shown (high effort or high expertise), the ratings of perceived quality are largely unaffected. For example, the mean rating in the High Effort / High Expertise condition (μ=3.53) is close to the mean of

Figure 3.4: The graph shows the mean ratings of quality across conditions (best in color). The x-axis clusters the conditions at the five levels of intrinsic quality. In each cluster, the left bar is the baseline and the bars are then ordered from the most negative (red) to the most positive (green) cues given about the provider. For the legend, L=Low, H=High, E=Effort, X=Expertise, NG=Not given. For example, HE / LX is the high effort / low expertise condition. Standard error = 0.023.



	df	SS	MS	F	p-value
Intrinsic quality	4	970.4	242.59	240.284	<0.0001**
Effort	2	89.3	44.64	44.214	<0.0001**
Expertise	2	96.7	48.37	47.909	<0.0001**
E:X	4	16.6	4.16	4.12	0.0025**
I:E	8	8.5	1.07	1.056	0.39
I:X	8	9.6	1.2	1.191	0.30
I:E:X	16	10.9	0.68	0.674	0.82
Residuals	2655	2680.5	1.01		

Table 3.1: Summary of three-way ANOVA applied to the perceived quality ratings. ** = significance at 0.05.

the baselines (μ=3.49, n.s.). Intrinsic quality also has a significant main effect on the ratings of perceived quality, as expected, but does not interact with the other two factors.

3.2.1 Study One: Discussion

The main results of the experiment are (i) showing cues of a provider's effort and expertise affect judgments of the quality of their feedback; (ii) negative cues has the largest effect on ratings of perceived quality, reducing ratings up to 21% relative to the baseline conditions, while positive cues has little impact; and (iii) effort cues are weighed similarly to expertise cues for judging feedback quality.

An interesting pattern in the results is that the negative cues serve to reduce ratings of feedback quality without a commensurate increase in ratings for the positive cues. This pattern is consistent with the "negativity bias" in psychology showing that negative information influences evaluations more than positive information [133]. The lack of influence of the positive cues may be due in part to participants assuming a certain level of effort and expertise on behalf of providers in the baseline conditions. Our results are inconsistent with other work showing that positive cues lead to more favorable impressions [50, 131]. One explanation could be due to different evaluation targets. In our study, the target is the feedback responses, whereas in [50, 131], the evaluation targets were other people. Positive cues may have weaker effects because it is easier to evaluate the quality of the feedback than to evaluate people's innate talents. The differences could also be due to the presentation style, granularity, and the number of cues provided alongside the evaluation targets.

Affective priming theory potentially offers an alternative explanation of our results [38]. Signaling low effort or expertise could be considered a negative prime. We are skeptical of this explanation for two reasons. First, the cues about the provider (the prime) were read after reviewing the feedback, whereas priming typically requires seeing this information first. Second, if acting as an affective prime, one would expect to see increased ratings of quality for the positive cues, which were not present in the data collected. Additional aspects and implications of our study will be discussed in the General Discussion.

In our study, manipulating the effort and expertise cues was straightforward. But how could these cues be determined in a real-world crowdsourcing or other platform where the feedback exchange is typically remote and anonymous? For expertise, system designers could apply known techniques such as performance-based assessments [134], aptitude tests [135, 136], or peer prediction [137]. For effort, however, there has been little research aimed at measuring it in a crowd context. Solutions such as self-reports may be ineffective due to strong biases against negative self-assessment (e.g. would workers really report they made

little effort on a task?), especially if linked to negative outcomes such as having the work rejected on a paid platform [124].

We therefore report on a second study which tests whether recording behavioral data collected during a design feedback task could be leveraged to predict the overall effort perceived by human raters. The study also contributes a new method for judging effort in an experimental setting.

3.3 STUDY TWO: METHODOLOGY

In study two, we explore the use of behavioral data for modeling perceived effort for design feedback tasks in a crowdsourcing context. The approach was to first collect behavioral data from providers leaving feedback for three designs and independent ratings of their perceived effort. We then built statistical models to learn mappings from features derived from the behavioral data to the ratings.

3.3.1 Behavioral Data Set

To collect behavioral data, we instrumented the feedback collection interface described in Study 1. See Figure 3.1. The interface was therefore used to collect both the feedback and the behavioral data for this study. Feedback and the associated behavioral data was collected for three Web designs; the home page of a community college (shown in Figure 3.1), an event organization site (http://evite.com), and a site for disseminating recorded talks (http://ted.com). The latter two sites along with some of the feedback provided are shown in Figure 3.6.

Using scripts added to the collection interface, we recorded the task behavior of the provider including mouse activity, keystrokes, interface and window actions, and start and end times for the main parts of the task. All events were time stamped. The scripts did not interfere with performing the task. Providers were not aware of this data collection.

The feedback collection interface was developed to aid the timings. For instance, after reading the general instructions, the provider had to select a button to reveal the design image and begin leaving feedback. This allowed us to record the preparation time (time from the onset of the task to the reveal of the image) and the design review time (from the reveal of the image to the first action). Sixty feedback responses were collected for each design, giving 180 total.

3.3.2 Replay Tool

Effort is how hard a provider works to give feedback on a design (e.g. how long did s/he view the design) and needs to be judged based on his or her behavior rather than on the content of the response. For instance, if a provider gave useful feedback but the content was blindly pasted from another source, then their effort on the task was minimal.

To enable effective judgments of effort, we built a tool that read the behavioral data and replayed (in the form of a video) the provider performing the task. To minimize influence from the feedback content, each character entered in the edit box was replaced with an 'x'. The purpose was to focus the judges' attention on the behavior (e.g. typing speed and content revision), rather than the content itself. If the idle time between actions was longer than five seconds, the tool enabled a "skip" button for jumping to the next recorded action. If the tool found a top-level window focus event during the idle time, it displayed a message that the provider switched to another window.

3.3.3 Judges

Three graduate students from our institution were recruited to judge the effort made by each feedback provider using the replay tool. The students were not affiliated with the project and did not participate in Study 1. Performing the ratings took about three hours and each judge was paid $85.

3.3.4 Procedure

After informed consent, judges received an overview of the study along with a description of effort - how much energy the provider invested in providing the feedback. Judges were presented with a sample of the replays to calibrate their ratings. The researchers informed the judges that they were free to develop their own criteria for judging the effort observed but suggested considering aspects of the entered text, annotation, and duration. The judges rated the effort on a scale from 1 (low effort) to 5 (high effort). Ratings could only be made at the end of a replay and were entered into a spreadsheet shown on a second monitor. Judges were allowed to revisit replays and modify their ratings until satisfied. Each judge rated the effort of the 180 providers who gave feedback. The three ratings of each provider were then averaged to produce the final rating. Though the size of the data set was modest, it was sufficient for testing the feasibility of mapping the behavioral data to the ratings.

	Features	Explanation	Gain (rank)
Annotation	Strokes	Number of strokes used for the annotation	0.12 (17)
	Stroke colors	Number of colors used for the annotation	0.18 (9)
	Bounding box area	Area of bounding box for the annotation	0.29 (7)
	Bounding box percent	Size of bounding box relative to the design image	0.28 (8)
	Average speed	Average cursor speed	0.14 (14)
	Max speed	Maximum cursor speed	0.15 (12)
	Average acceleration	Average cursor acceleration	0.12 (18)
	Max acceleration	Maximum cursor acceleration	0.13 (15)
	Covered area	Percent of pixels in bounding box covered by the annotation	0.18 (10)
	Overpaints	Pixels painted by multiple strokes	0.06 (20)
	Undos	Number of stroke undos	0
Text Entry	Control actions	Total number of actions on the annotation control panel	0.07 (19)
	Pauses	Number of pauses longer than two seconds when entering text	0.36 (5)
	Deletions	Number of deletions during text entry	0.37 (4)
	Characters	Character count of the text	0.59 (2)
	Words	Word count of the text	0,64 (1)
	Average word length	Average word length of the text	0.17 (11)
	Longest word	Length of the longest word in the text	0.25 (9)
	Typing speed	Average speed for typing the text	0.13 (16)
	Insertions	Number of char insertions	0.14 (13)
	Text ratio	Ratio of total annotation time to total time entering text	0.34 (6)
Timings	Task time	Total time spent on the task	0.40 (3)
	Prepare time	Time taken to read the general description (from start of task to selection of "start feedback")	0
	Image review	Time taken to review the design (from selection of "start feedback" until first action).	0
	Task review	Time from last action until the task is submitted.	0

Table 3.2: The features used for modeling perceived effort. Right column shows the information gain scores (rank) for each feature for the binary classification.

3.3.5 Features

We created 25 features from the behavioral data and these are shown in Table 3.2. The features were derived from discussions with the judges about what observed behaviors affected their ratings, our experience piloting the tasks and data collection, and prior work [24, 138]. The features are not exhaustive, but do provide a reasonable starting point for exploring statistical models of perceived effort. A feature vector was created for each feedback response.

3.4 STUDY TWO: RESULTS

The rating distribution of the judges is shown in Figure 3.7 and was nearly uniform (μ=3.0). This validates that the feedback providers (workers) performed the feedback tasks with varying levels of effort. Krippendorff's alpha was 0.79, indicating good agreement among the judges. The fact that judges could agree on the effort observed suggests that statistical models could also learn the mappings.

All models were built using support vector machines in Weka 3.6 and tested using ten-fold cross validation. Alternative statistical models including logistic regression, naive Bayes, and decision trees were also explored. These models produced similar results to what is reported below.

As a first step, we created models that learned mappings from the features to the five levels of effort. The results are summarized as a confusion matrix in Table 3.3 (left). The overall accuracy was 65%, precision was 0.65, recall was 0.64, and the F-measure was 0.64. From the table, the most egregious errors (e.g. actual low effort predicted as high effort or vice versa.) were rare. The accuracy was modest, but may be improved by training on a larger data set and extracting additional features from the behavioral data.

As an alternative to predicting the five levels of effort, we simplified the problem to a binary classification; effortful (ratings of 3, 4 or 5) and not effortful (ratings of 1 or 2). A binary classification would be easier to interpret and would be consistent with the two levels of effort manipulated in Study 1. A model was trained using the same data, but now for the binary classification. The accuracy was markedly improved (92%). Precision, recall, and the F-measure were all 0.92 and the results are shown in Table 3.3 (right). To determine which of the features contributed most to the classification, we performed feature selection using the information gain metric in Weka. The gain scores for each feature and their rank are shown in the right column of Table 3.2. The length of the text, total time on task, revisions made to the text, typing pauses to (presumably) review the design, and multiple colors in

		Predicted Rating				
Actual Rating		**1**	**2**	**3**	**4**	**5**
	1	21	7	0	1	0
	2	3	30	10	0	0
	3	0	11	20	4	1
	4	0	4	10	18	7
	5	0	0	0	7	26

	1	**2**
1	91	6
2	8	75

Table 3.3: The confusion matrices for predicting effort for the five levels (left) and for two levels (right). 1=low effort (both), 5=high effort (left), 2=effortful (right).

the annotation were among the features that contributed most to the classification.

3.5 DISCUSSION

The results from Study 2 showed that it is feasible to model perceived effort for crowdsourced design feedback with up to 92% accuracy. There are at least two ways that crowd feedback services could apply this finding. One way is for the service to model the perceived effort of the providers and display the classifications (cues) for the designer. From the results in Study 1, services may only need to show the cues when they would be most beneficial for differentiating the feedback. A second way would be for the service to use the models to automatically reject low effort work and acquire more effortful responses. This approach would trade feedback generation time for a set of responses that are likely to be of higher quality. In fact, there was a strong positive correlation (Pearson's $r = 0.82$, $p<0.001$) between the ratings of perceived effort and quality for the feedback data set shared by the two studies in this chapter.

We modeled perceived effort for crowdsourced design feedback, but the approach generalizes to other tasks where users must judge subjective responses from the crowd. Such tasks can be found in crowd-based ideation [122], content summarization [139], and social data analysis [127]. Leveraging our methodology from Study 2, including the behavioral data collection and replay tools, researchers and system designers can build statistical models for their own tasks. It may also be possible to build a more general model by considering only lower-level features independent of the task type and training it on a larger data set [140].

Logging behavioral data to make effort visible in a crowd service could raise privacy concerns. However, it could also lead to practices favorable for workers. For example, crowd services could enable users to pay bonuses based on the effort invested by workers. Workers may also improve their performance if they are able to view and reflect on their own effort, and possibly command higher pay with a reputation of effortful responses. In addition to

showing cues that summarize effort, a service might also show how the worker's logged behavior compares to other workers for the same task. This could be used to explain the cues or to improve performance by showing replays of effortful work as exemplars.

The results from Study 1 showed that expertise cues also factor into judgments of feedback quality. Expertise measures could be implemented as qualification or screening tasks that gauge relevant aptitudes [135, 136], measure peer prediction ability [137] or apply performance-based assessments [134]. Future work could also explore extending models of user interface skill (e.g. [141, 142]) to model the domain expertise of crowd workers.

The manipulations of effort and expertise cues in Study 1 were achieved using specific phrasings of text. Researchers have already shown that different representations can have different influences over the evaluation of work quality [121]. This thread of research could therefore be extended to study different phrasings and granularities of the cues used in our study and in context of design feedback. It would be also interesting to study whether the cues always need to be displayed or only when needed to differentiate the feedback.

In this preliminary study, we find that negative information about the feedback providers has a stronger impact over people's perception of the feedback than positive information does. This finding piqued my interests in how feedback written in a negative tone affects the recipient's revisions to the project and their perceptions of the provider and the feedback. The written content itself serves as a cue that is interpreted by the feedback recipient and this cue may be as important as explicit cues of effort and expertise. My main thesis work focuses on studying the impacts of these negative feedback and propose interventions that address the problem.

In the next three chapters, we describe the proposed interventions and report experimental results. First, we examine three coping activities, namely self-affirmation, expressive writing, and distraction. Recruited content creators perform these activities before revising their work in an online study. Next, we explore a feedback reordering mechanism that presents the content based on feedback valence. Lastly, we test whether empathy arousal could motivate participants to be more supportive and offer higher quality feedback. The study evaluates both a narrative based and an ingroup framing based approach.

Figure 3.5: A graph of the perceived quality ratings collapsed across intrinsic quality. The legend is the same as Figure 3.4

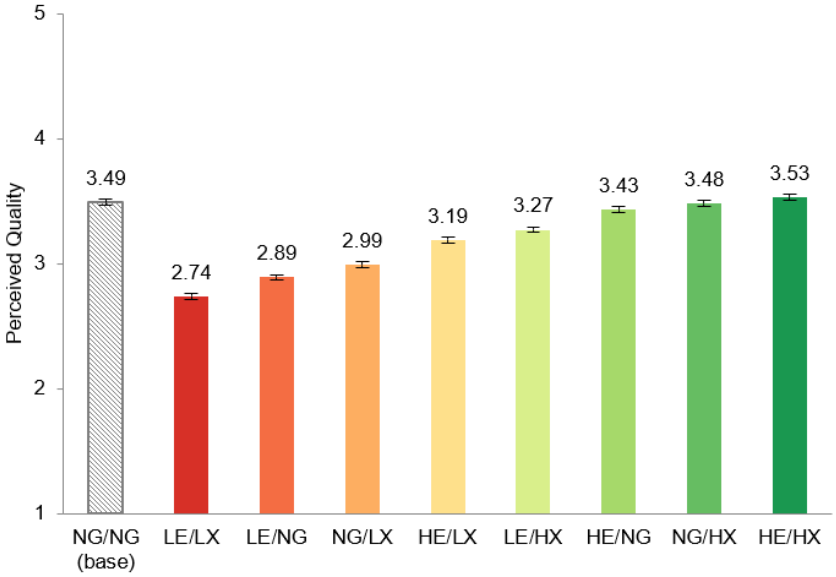


Figure 3.6: The two additional Web designs for which crowd feedback and behavioral data was collected for modeling perceived effort. The task instructions and annotation panel were omitted here but were the same as shown in Figure 3.1

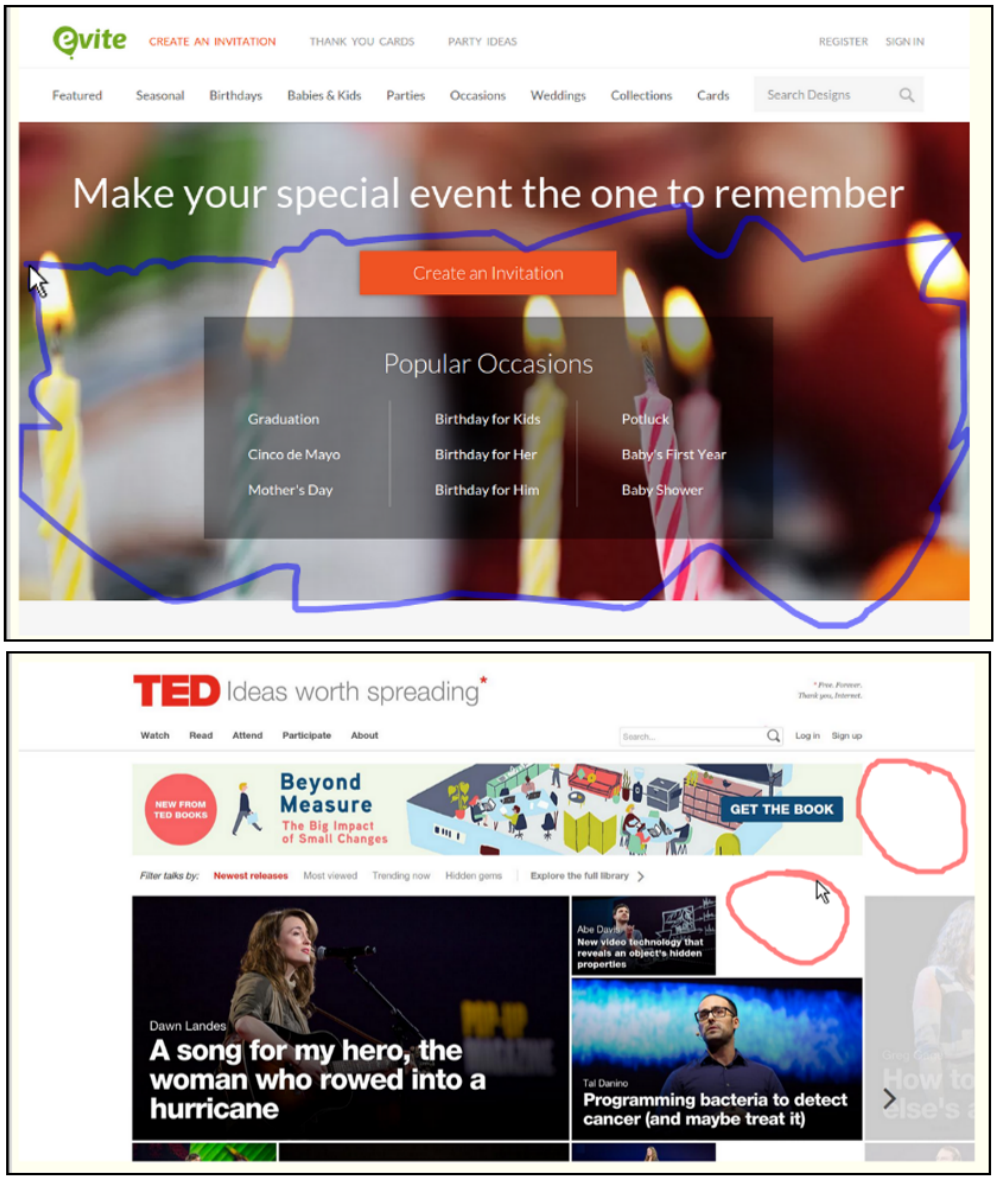
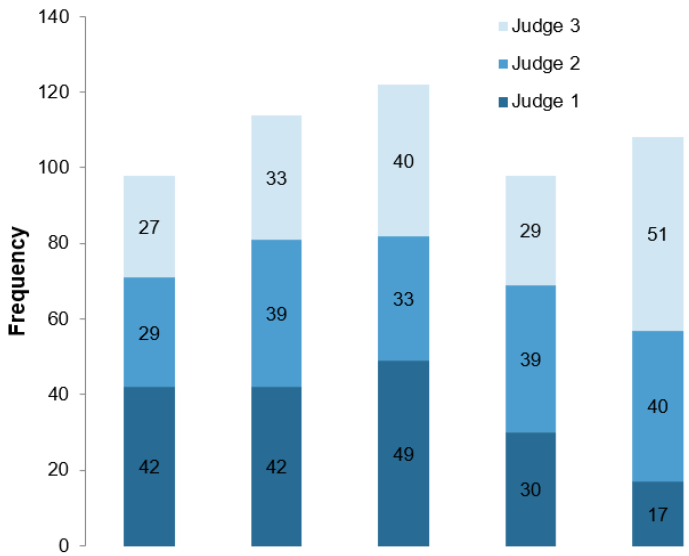


Figure 3.7: The distribution of ratings of the perceived effort invested by providers for the feedback tasks. There was good agreement overall and the distribution was close to uniform.



CHAPTER 4: COPING ACTIVITIES

In this chapter, we examined three coping activities which we hypothesize may mitigate the effects of negative feedback: self-affirmation, expressive writing, and distraction. All three coping activities have the potential to increase people's resilience to negative feedback as suggested by prior work. Self-affirmation activities inhibit defensive reactions to ego-threatening information, such as negative feedback, by affirming people's core values [30, 31]. The affirmation could uphold people's self-worth and encourages feedback reception. Expressive writing facilitates recognition and expression of stress-related thoughts and negative emotions [32]. By reducing the interference from distress, expressive writing also improves performance in cognitively loaded tasks, such as test-taking [33]. In our study, we hypothesize that expressive writing can lessen the distress caused by the negative feedback and increase participants' performance on the experimental task. Distraction relieves distress and anxiety by directing a user's attention away from the source [34, 35]. A short duration of mind wandering has been shown to improve people's affective states and also stimulates the development of novel solutions to previously encountered problems [36]. After receiving negative feedback, distraction may help content creators to recover from distressful affective states, interpret the feedback from a new perspective, and conduct more effective revision.

Researchers have explored additional coping activities but many of these activities are not suitable for online environments, such as physical exercise [143], meditation [144], spirituality and religion [145]. In our work, the three selected activities have pathways to be implemented as standalone interventions compatible with existing feedback collection workflows. Besides comparing these activities in the context of iterative design process, our work is also original because we study the activities with feedback sets containing different balances of negative valences, such as mainly negative or all neutral. Prior work has examined the activities when participants receive stark negative valence information from a single source [30, 31, 32, 34, 35]. But in online environments, content creators usually receive a set of feedback with mixed valences from multiple sources. Here we simulated this setup and evaluated the activities in a realistic setting.

To compare the three coping activities, we conducted a full factorial experiment with coping activities and valance balances as factors. Each feedback set had three pieces of feedback, where the valence balance ranges from all negative to all neutral. The experiment included two phases. In the first phase, participants wrote an essay on a complex social issue. In the second phase, participants performed one of the coping activities and revised

their essay based on a provided feedback set. All participants received feedback referencing the same aspects of their essay but with different valence balances based on experimental condition. We measured participants' affective states and extents of revision to quantify the impact of negative feedback and the coping activities. Following prior work [27, 146], we also measured participants' perception of the feedback and its providers as they relate to the receptivity of the feedback.

Our results showed that receiving a feedback set containing one piece of negative feedback significantly raised participants' ratings of negative affects by 55%, reduced ratings of positive affects by 15%, reduced the extent of the revision by 28%, and lowered the perception of feedback and its providers by 24%. This result highlights that even a small amount of negative feedback can have a notable impact. Among the three activities tested, expressive writing encouraged essay revision while distraction improved participants' affective states and their perception of the feedback providers. Self-affirmation had no significant effects. Our results showed that no single activity outperformed the others. Platform designers could choose which activity to use based on how the designers prioritize different measures or situational needs. Future work can build upon our results and explore other activities to offer more coping methods within this emergent framework.

The HCI contributions of this experiment are (i) empirical knowledge of how feedback sets with different valence balances impact users' affective states, revision behaviors, and perceptions for a writing task; (ii) deeper empirical understanding of how three theoretically-based coping activities mitigate the effects of negative feedback; and (iii) practical guidelines regarding when to use the coping activities to improve users' resilience to negative feedback online: using expressive writing when valuing revision behavior the most and using distraction when prioritizing users' affective states or perceptions of the feedback providers.

4.1 METHODOLOGY

We first investigate activities that mitigate the influences of negative feedback. Specifically, we are interested in evaluating three theory-based coping activities: self-affirmation, expressive writing, and distraction. In addition, we test the activities with feedback sets of different valence balances to have a more realistic and more fine-grained evaluation of the effects. In the second part of my thesis, we focus on answering two research questions:

- RQ1: How do feedback sets with different balances of valence affect participants' affective states, extents of revision, and perceptions of the feedback and its providers?

- RQ2: To what degree can coping activities based on theories of self-affirmation, expressive writing, and distraction, mitigate the influence of negative feedback on these same measures?

Answers to these questions will deepen empirical knowledge about the effects of receiving negative feedback and how to reduce those effects. Answers will also provide insights regarding practical interventions that improve user's resilience to negative feedback received online. In this section, we describe how we conduct the study and what measurements are collected.

4.1.1 Experimental Design

To answer the research questions, we conduct a full-factorial between-subjects experiment with two factors: coping activity and valence balance. Coping activity examines four interventions: self-affirmation, expressive writing, distraction, and a control (no activity). Valence balance refers to the number of pieces of feedback in a set with a neutral/negative orientation. This factor had four levels: all neutral (receiving three pieces of negative feedback), mainly neutral (receiving two pieces of neutral feedback and one negative), mainly negative (two negative, one neutral), and all negative.

4.1.2 Task Setup

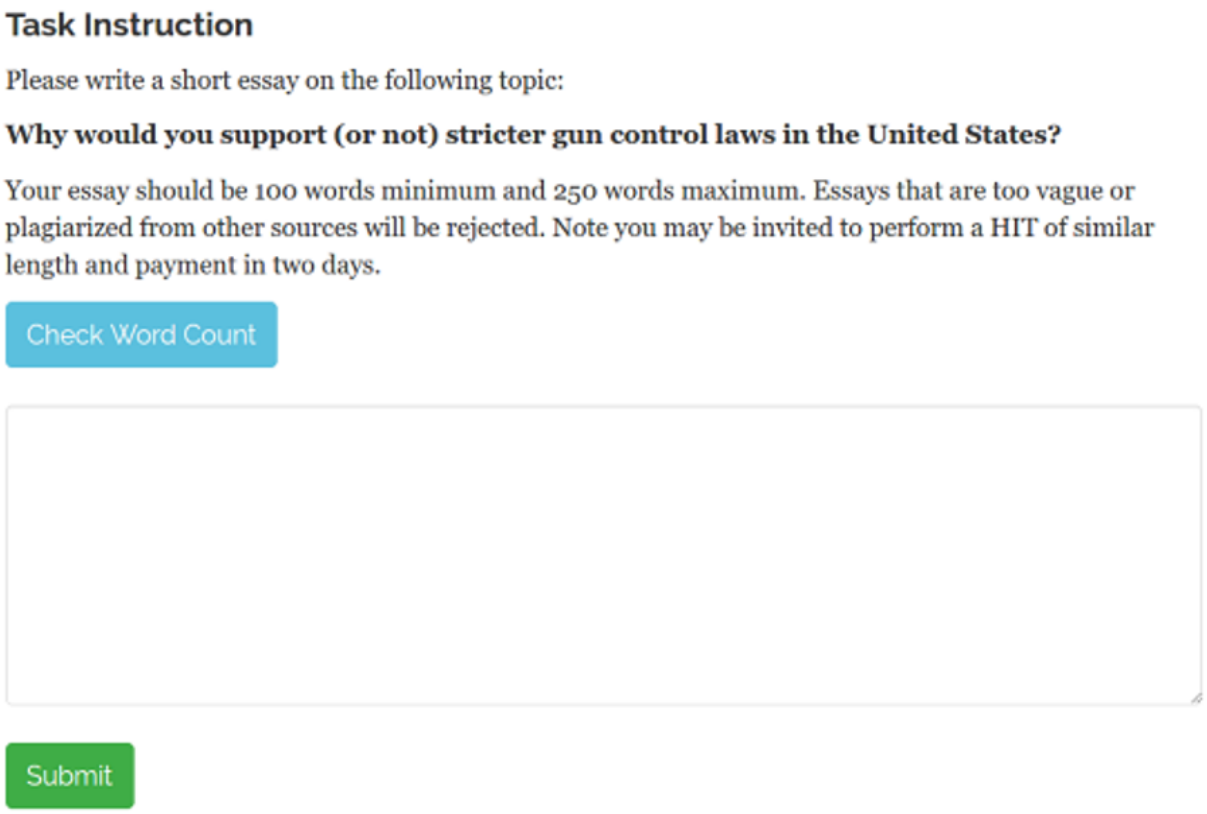
The experimental task is an essay composition task including a writing phase and a revision phase. In the writing phase, participants write an essay about whether they would support stricter gun control laws in the U.S. This topic is selected because it is widely debated and familiar to a general audience in the U.S. In addition, participants may have an existing stance on the topic and genuinely care about the feedback. The task instructions state that vague or plagiarized essays will not receive payment. We enforce a 100-250 word limit and a 30-minute time limit so each participant invests similar amount of effort in the task. Participants could track word counts by selecting a button on the task interface. In the revision phase, participants receive three pieces of feedback and revise their essays in a text box pre-filled with the content. Participants have 30 minutes to finish this phase.

4.1.3 Coping Activity Factor

Three coping activities are tested: self-affirmation, expressive writing, distraction, and control. See Figure 4.1-4 for screenshots of the intervention interfaces. All activities happen

in the revision phase. As prior work has not evaluated the efficacy of these activities in the context of receiving design feedback, here we test each activity individually to isolate the effects and leave the study of synergies of different activities for future work.

Figure 4.1: Screenshot of the writing phase of the experimental task.



In the self-affirmation condition, a participant reflects on the positive aspects of oneself. Following [147], a participant ranks six core values (business; art, music, and theater; social life and relationships; science and pursuit of knowledge; religion and morality; and government and politics) by how important these values are to him or her. After the ranking, a participant explains the importance of the top-ranked value. Prior work reports timing is crucial for the efficacy of self-affirmation, which needs to happen before participants face the ego-threat [30]. In our experiment, participants perform the intervention before reviewing the feedback.

In the expressive writing condition, a participant reviews the feedback, reflects on his emotional reactions, and expresses them in a provided text box. The task interface displays the feedback for reference. The instructions are adapted from prior work on emotional coping [148].

Figure 4.2: Screenshot of the self-affirmation phase of the experimental task.

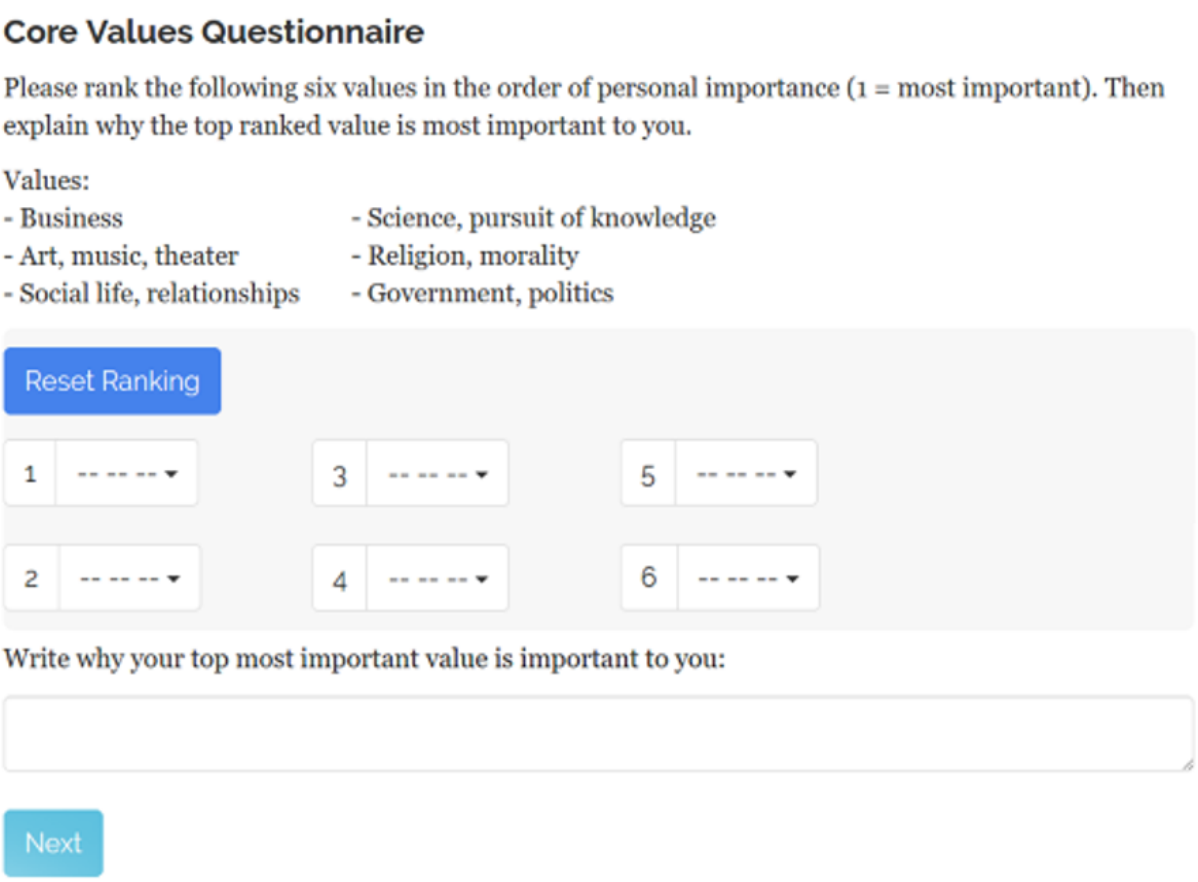


Figure 4.3: Screenshot of the expressive writing phase of the experimental task.

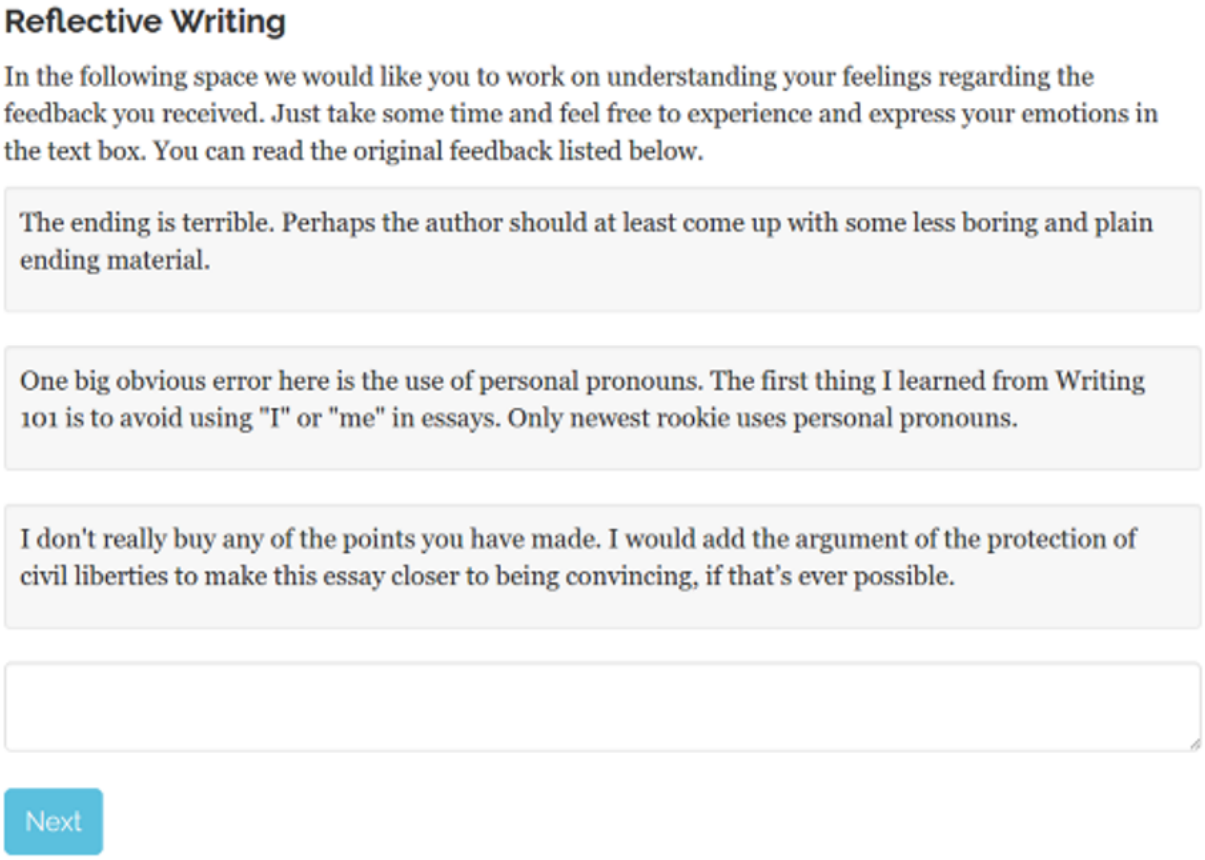
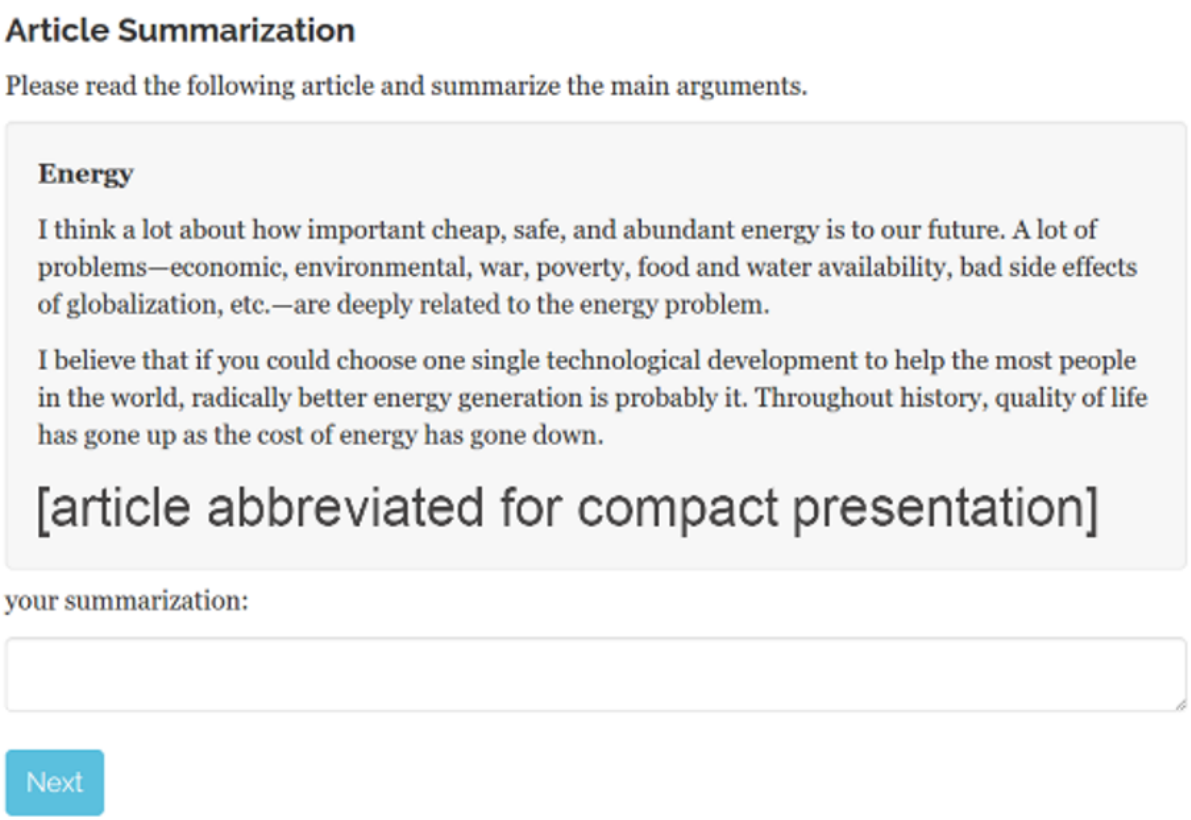


Figure 4.4: Screenshots of the distraction phase of the experimental task.



In the distraction condition, the participants are instructed to perform a reading comprehension task to divert their attention away from the feedback reflection. Participants read an article about energy consumption, an issue orthogonal to the essay topic, and summarize it in a text box. We conduct a pilot study to select an article with appropriate length that requires similar time to complete relative to the other conditions. The distraction intervention happens after feedback review and before the revision.

In the control condition, participants receive feedback and revise their essays, but do not perform any coping activity.

4.1.4 Valence Balance Factor

Valence balance has four levels: all negative (three pieces of negative feedback; labeled as "3-" in the tables and figures), mainly negative (two negative; labeled as "2-"), mainly neutral (one negative; labeled as "1-"), and all neutral (zero negative; labeled as "0-"). All participants receive feedback on the same aspects of the essay but phrased with different valence balances. In this way, only the feedback valence, rather than its content, differs in the experiment. There are in total three pairs of feedback (Table 4.1). Within each pair, the only difference between the two pieces of feedback is the phrasing of the content: one neutral and one negative. During the experiment, each participant receives one piece of feedback from each pair and three pieces of feedback in total. In case participants do not find the provided feedback useful, we allow them to reject any piece of feedback and revise the essay in any way they deem appropriate.

The feedback delivered to participants is derived from authentic feedback compiled online. We first collect five essays from Amazon Mechanical Turk (AMT) and use them to solicit a large feedback pool on three core aspects of the essays: content, structure, and style [149]. For each aspect, we select the piece of feedback that is the most generalizable and actionable, and has neutral valence. To ensure the selected feedback is applicable to the essays, we use a script to filter out incompatible essays during the experiment. Since many people have strong beliefs about gun control policy, we only chose feedback with a neutral stance and a sole focus on the essay quality. We correct misspellings and grammatical errors in the selected feedback to prevent language bias. For the feedback selected for each aspect, we create a piece of complementary feedback by modifying its language to be more negative. This forms a neutral/negative pair in the final feedback set. The final set has one pair of feedback for each of the three aspects. In the essay revision phase, each participant receives three pieces of feedback covering all core aspects. For example, a participant in the mainly negative condition may receive one piece of negative feedback on the content aspect, one

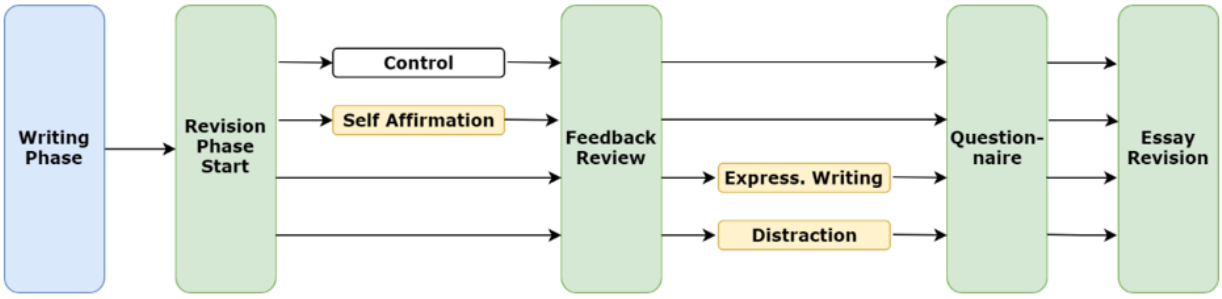
negative on structure, and one neutral on style. This setup can mitigate any confounding effects caused by participants being more receptive to feedback on specific aspects.

Three graduate research assistants not affiliated with this project review the final feedback set. They report no difference in the valence among the negative feedback and among the neutral feedback. They also report that the negative feedback has a notably more negative valence than the neutral feedback.

4.1.5 Participants

We recruit 681 participants in total, among which 518 participants finish both phases. 38 excess data points collect at the end were excluded. Given the scale of the experiment, the recruitment took place on AMT. We configure the task to require all participants to reside in the U.S. given the topic of the essay and to mitigate issues of language proficiency. 77% of participants complete an optional demographic survey. Among these participants, 50% report their gender as female and nearly all (99%) selected English as their first language. For age, 81% report being between 18-44 years of age and 19% report being 45 years of age or older. The highest level of education is reported as high school (41%) and an undergraduate degree or higher (58%).

Figure 4.5: Graphical summary of the procedure.

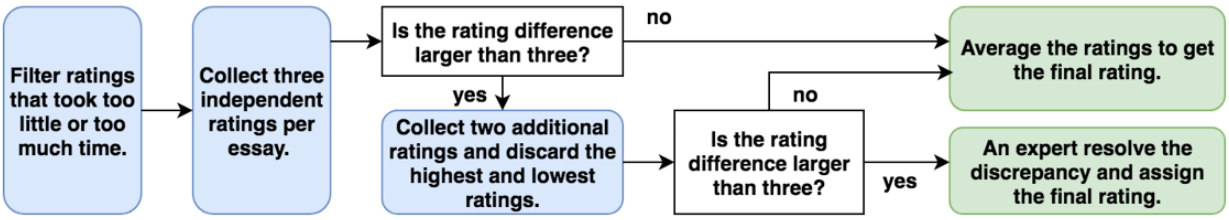


4.1.6 Procedure

Figure 4.5 shows the experiment workflow. At the start of the writing phase, participants sign an IRB consent form and read an overview of the workflow. After they complete the writing phase, a script filters out 3% of the essays that do not use personal pronouns or keywords related to civil liberties, as it would be inconsistent with the feedback on style or content (Table 4.1). Since participants may question the viability of receiving feedback

immediately after they finished the essay, we instrument a two-day delay to simulate realistic feedback collection from an online platform. The delay also mitigates potential confounding effects caused by valence arousal during the writing phase. In the revision phase, we assign participants a set of feedback with a valence balance determined by the assigned experiment condition. Participants complete a survey (see Measurement) immediately after performing the coping activity. We reward participants $2 for completing the writing phase, and an additional $3 for completing the revision phase. The payment rate is determined by a pilot study to be consistent with U.S. minimum wage. A follow-up survey asks participants about how important the gun control issue is for them and how frequently they write on this topic. The collected ratings are used as covariates in the analysis. After the study, participants are debriefed via email.

Figure 4.6: Graphical summary of the essay rating procedure.



4.1.7 Measurement

We measured three categories of dependent variables: participants' affective states and perceptions, extents of the revision, and behavioral data from both task phases. For the affective states and perceptions, we collected the measurements via a survey including eleven questions:

- Four items regarding how happy / enthusiastic / annoyed / frustrated the participant feels. The questions were adapted from PANAS [150].
- Three items regarding how positive / useful / fair the feedback is.
- Four items regarding how considerate / polite / knowledgeable / (exhibiting) expertise the feedback providers are. The items were adapted from previous work on feedback reception [27].

For each statement, participants rated their degree of agreement on a seven-point scale (1=Strongly Disagree, 7=Strongly Agree).

For behavioral data, a script counted the number of characters edited during revision and calculated the final edit distance from the original essay [151]. We also asked participants to report how many and which piece of feedback they incorporated into their essay. We also logged the time participants spent composing essays, reviewing feedback, performing coping activities, and revising essays.

Given the scale of the data, we recruited 288 judges from MTurk to rate the quality of the initial and revised versions of the essays. We provided rubrics defining the three core aspects of the essays, namely content, style, and structure, together with examples for each aspect. Following the rubrics, the judges evaluated the quality of the essays on a 7-point scale (7=high quality). To calibrate the rating scales, each judge rated a set of 10 essays randomly assigned by a script. At the end of the rating session, the judges could adjust their ratings on a page displaying the essays and their ratings.

We discarded 4.4% of the ratings which took the judges too little time (less than three seconds) or too much time (two standard deviations above the mean) to assign. In the end, each essay received ratings from at least three independent judges. If the judges reached a consensus where the maximum difference among ratings was fewer than or equal to three units, we averaged the ratings to produce the final rating; if no consensus was reached, we collected two additional ratings and discarded the highest and lowest ratings. If the discrepancy remained, an expert in writing was recruited from Upwork to assign the final rating. Overall, the judges reached a consensus in the first round for 74.6% of the essays and the expert resolved the discrepancy for 3.1% of the essays. Figure 4.6 shows the essay evaluation workflow.

4.2 RESULTS

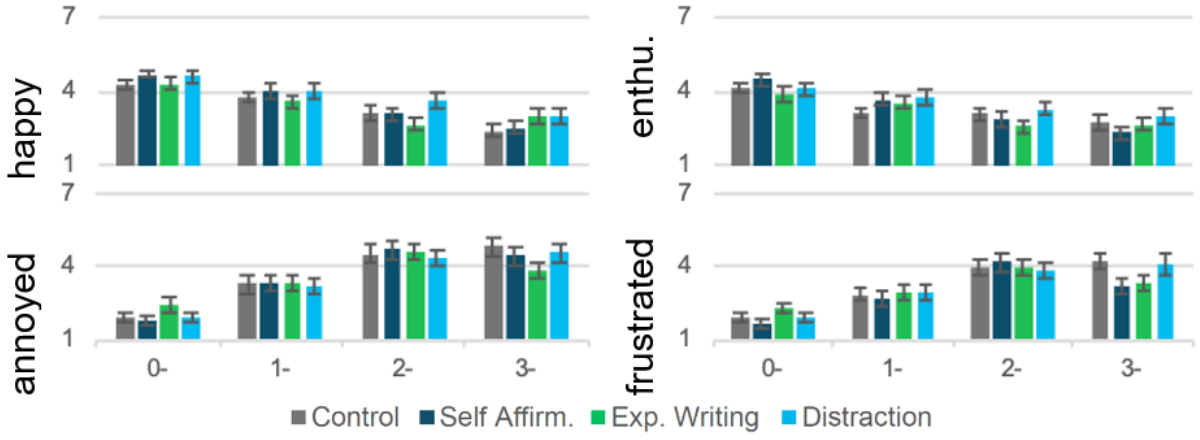
We report how valence balance and coping activity influence participants' affective states, extents of revision, and their perceptions of the feedback and its providers. Participants find the topic of gun control moderately important with a rating of 3.97 (SE=0.11) out of 7. The essays have an average word count of 171.6 (SE=10.5), which is substantially higher than the 100-word minimum limit. In this section, we focus on reporting statistically significant results and highlight patterns of interest for follow-up discussion. Using how much participants care about gun control and how frequent they wrote on this topic as covariates did not change significance levels. Therefore, we report the results of the analysis without using covariates. See Table 4.2 for all ANOVA results.

4.2.1 Affective States

A MANOVA analysis shows valence balance has a main effect on participants' affective states ($F[12, 1389]=1.22$, $p<.001$). In comparison with the all neutral condition, increments in the feedback negativity significantly reduces participants' affective states ratings until the feedback set becomes mainly negative (Table 4.3). An ANOVA shows coping activity has a marginal effect on the happiness rating ($F[3, 468]=2.20$, $p=.088$). No significant interaction effect is detected between coping activity and valence balance.

Among the coping activities, distraction is the only activity that has significant influences over participants' ratings of affective states. The distraction intervention significantly increases participants' happiness rating (M=3.81, SE=0.15) in comparison with the control condition (M=3.4, SE=0.14). The other two activities have no significant influences over the affective states, but we do observe some trends consistent with prior work. In the all negative condition, expressive writing and self-affirmation tend to improve participants' affective states (Figure 4.7). When all three pieces of feedback are negative, participants in the expressive writing condition report being happier (M=3.0, SE=0.28), less annoyed (M=3.8, SE=0.32), and less frustrated (M=3.27, SE=0.33) than the control condition (M=2.37, SE=0.25; M=4.83, SE=0.36; M=4.23, SE=0.37), an average difference of 0.87 units on the measurement scale. Participants in the self-affirmation condition rate their frustration lower (M=3.23, SE=0.32) than the control condition (M=4.23, SE=0.37).

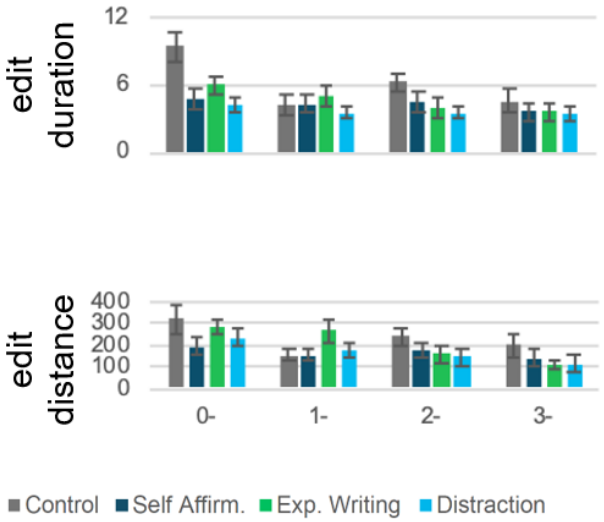
Figure 4.7: Bar charts for affective states. Distraction increased happiness rating. The y-axes refer to the 7-point Likert scale rating.



4.2.2 Essay Revision

Valence balance has main effects on the edit distance ($F[3, 468]=5.80$, $p<.001$) and edit duration ($F[3, 468]=5.71$, $p<.001$). Participants in the all neutral condition spend significantly more time and edit the essay to significantly greater extent than the participants in the other valence conditions do. In comparison with the all neutral condition, edit duration decreases by 38.3% and edit distance decreases by 45.7% in the all negative condition (Table 4.4). Participants receiving more negative balanced feedback spend less time in revision and edit the essays less (Figure 4.8).

Figure 4.8: Bar charts for extents of revision. Participants in the control condition performed no additional activities and edit the essays more. Expressive writing encourages revision in comparison with the other two coping activities.



Coping activity has a main effect on the edit duration ($F[3, 468]=6.01$, $p<.001$), and a marginal effect on the edit distance ($F[3, 468]=2.14$, $p=.094$). Participants in the control condition spend significantly more time editing essays than the other three activity conditions (Table 4.4). Also, participants edit significantly more characters in the control condition than in the self-affirmation and distraction conditions. In comparison, participants in the expressive writing condition report similar level of edit distance as in the control condition.

Participants receiving negative feedback edit their essays to half the extent that participants did when receiving the same feedback written in a neutral tone. Regarding the coping activities, participants in the control condition spend more time on the revision task. This may be caused by the fact participants receive the same payment despite needing to perform additional work in the coping activity conditions. Notably, expressive writing leads to same

amount of essay revision even after participants write a short essay on their emotions.

4.2.3 Perception of Feedback Set

A MANOVA analysis shows valence balance has a main effect on participants' perception of the feedback set ($F[9, 1392]=28.68$, $p<.001$). Participants perceive negative feedback significantly less fair, less useful, and less positive compared to the all neutral condition (Table 4.5). Showing an additional piece of negative feedback significantly lowers participants' perception of the feedback set in all valence balance conditions.

Coping activity has a main effect on the usefulness rating ($F[3, 468]=3.56$, $p=.014$). In the expressive writing condition, feedback is rated significantly less useful than in the control condition (Table 4.5). Overall, self-affirmation has no significant effects. But in the all negative condition (Figure 4.9), we do observe self-affirmation notably lowers the ratings of positivity (M=1.33, SE=0.12), usefulness (M=2.73, SE=0.30), and fairness (M=2.87, SE=0.29) in comparison with the control condition (M=1.83, SE=0.27; M=3.90, SE=0.39; M=3.77, SE=0.37), an average decrease of 0.86 units on the measurement scale.

4.2.4 Perception of Feedback Providers

A MANOVA analysis shows valence balance has a main effect on the ratings about feedback providers ($F[12, 1389]=27.05$, $p<.001$). Participants give feedback providers significantly less favorable ratings as the negativity in the feedback set increases (Table 4.6). Similar to the trend observed in feedback perception ratings, showing one more piece of negative feedback significantly lowers participants' perception of the feedback providers in all valence balance conditions.

Coping activity has main effects on the consideration ($F[3, 468]=3.32$, $p=.020$) and politeness ratings ($F[3, 468]=2.81$, $p=.039$). Participants in the distraction condition rate the feedback providers significantly more considerate and polite than in the self-affirmation and expressive writing conditions (Table 4.6). In the distraction condition, participants also tend to rate the providers as more considerate and polite than in the control condition. Self-affirmation has no significant effects. But similar to trends observed in previous sections, self-affirmation tends to lower the evaluation of the feedback providers in the all negative condition. In this condition, self-affirmed participants rate the providers less considerate (M=1.7, SE=0.19), less polite (M=1.5, SE=0.15), and less knowledgeable (M=2.43, SE=0.23) than in the control condition (M=2.57, SE=0.31; M=2.2, SE=0.28; M=3.63, SE=0.31; see Figure 4.10), an average decrease of 0.92 units on the measurement scale.

Figure 4.9: Bar charts for participants' perception of the feedback set.

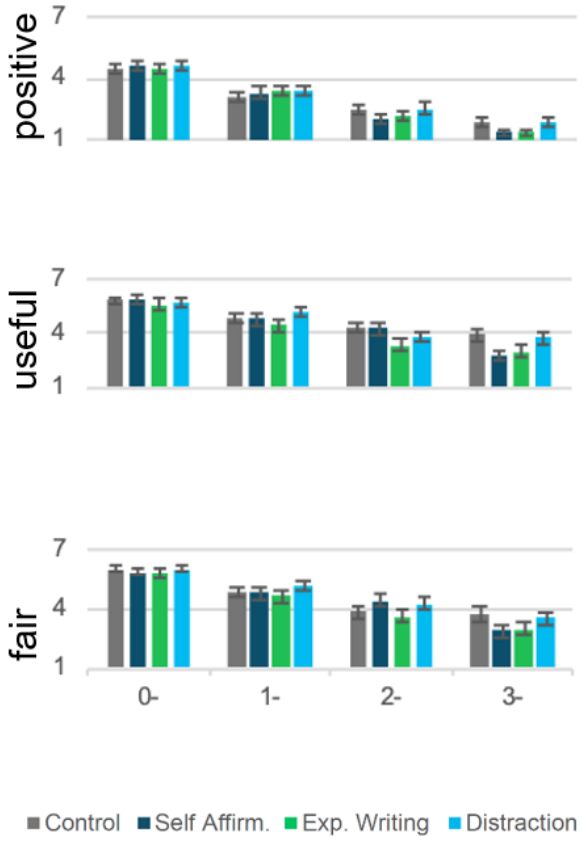
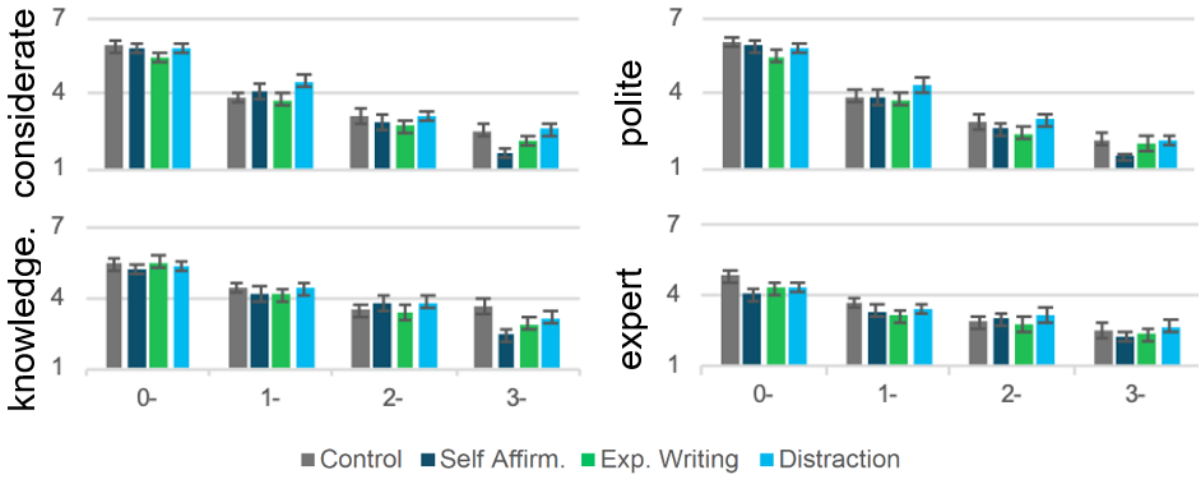


Figure 4.10: Bar charts for participants' perception of feedback providers. Distraction improves how considerate and polite participants perceived the providers to be in comparison with the other two coping activities.



4.2.5 Essay Quality

Revised essays (M=4.75, SE=0.09) receive significantly higher quality ratings than the original essays do (M=4.6, SE=0.09; $F[1, 958]=9.77$, $p=.002$). Neither valence balance nor coping activity has a main effect on the ratings. There is also no interaction effect between the factors and the essay version. Although valence balances and coping activities significantly impact participants' affective states and their perceptions of the feedback and its providers, these effects do not translate into higher quality essays after revision.

4.2.6 Accepted Feedback Count

Participants are more open to neutral feedback. Valence balance has a main effect on the number of feedback items reported to be accepted ($F[3, 468]=11.42$, $p<.001$). Participants report accepting more pieces of feedback in the all neutral condition (M=1.68; SE=0.09) than the all negative condition (M=0.93, SE=0.09; LSD=0.262, $p<.05$). Consistent with prior work, neutral feedback is more likely to be accepted than negative feedback [153]. Overall, 80% of neutral and 50% of negative feedback is accepted by participants. Coping activity has no statistically significant effect on this measure.

4.3 DISCUSSION

Our experiment shows that valence balance has significant effects on participants' affective states: the all negative valence balance condition reduces participants' ratings of their positive affects (happiness and enthusiasm) by 38% and raises the ratings of negative affects (annoyance and frustration) by 110% in comparison with the all neutral condition. Negative feedback significantly lowers the average rating of feedback being positive, useful, and fair by 49%, the average rating of the providers being considerate, polite, knowledgeable, and having expertise by 54%, and extents of revision by 46% relative to receiving all neutral feedback.

For the coping activities, self-affirmation has no statistical effects on the measures. However, in the all negative condition we observe self-affirmed participants have more positive affective states, which is a trend consistent with prior work [31, 70]. Expressive writing encourages significantly more revision in comparison with the other two activities. Distraction significantly improves participants' ratings of happiness and makes them perceive the feedback providers to be significantly more considerate and more polite compared to the other two coping activities.

Our results show various measures have different levels of sensitivity to negative feedback. Showing one additional piece of negative feedback significantly lowers participants' perception of the feedback and its providers in all valence balance conditions. For affective states, we observe a similar trend but there is no statistical difference between the mainly negative and the all negative conditions. For edit duration and edit distance, only participants in the all neutral condition report significantly higher measurements than the other valence balance conditions. This pattern of results indicates that platform designers should aim to eliminate any occurrence of negative feedback, as even one piece of negative feedback adversely affects edit duration, edit distance, and perceptions of the feedback content and its providers. For situations where users' affective states or their perceptions of the feedback are most important, such as when novice designers are collecting feedback, platform designers should focus on deterring negative feedback, especially snowballing effects [5]. Prior work has focused on examining whether presenting negative valence information lowers people's affective states and task performance in various contexts [6, 7, 8]. Our results show an equally important question is whether increasing negativity will continue to lower these measures. Even when we cannot eradicate the negative feedback, increasing the valence balance of the feedback set may still significantly improve the user experience.

While prior work suggested the selected coping activities could mitigate the influences of negative feedback over participants' affective states, our results indicate the activities are not as effective as we expected. Other than distraction, both expressive writing and self-affirmation had no significant effects over participants' affective states. One potential reason may be the mixed valence balance conditions we test in our experiment are more nuanced than the binary valence conditions in prior empirical studies, where participants receive either entirely negative or entirely non-negative information. While we do observe expressive writing and self-affirmation tend to improve participants' affective states in the all negative condition, a trend consistent with prior work [31, 70, 76, 77], the activities are not effective enough in the other valence balance conditions to show significant effects across conditions. Future work is needed to evaluate these activities with feedback sets that exhibit valence properties to further tune the associated theories. It is also possible that the activities are less effective because design feedback is more subjective than the negative valence information used in prior work, such as health-risk information [72]. Participants may view the provided negative feedback as a matter of opinion rather than a direct threat to their ego and thus do not benefit from the coping activities that address the potential threat.

In authentic settings, platform designers could offer users optional opportunities to perform the coping activities. In this way, only the users who feel the need for assistance

would perform the activities; while not affecting others. Our results show expressive writing encourages significantly more revision and distraction significantly improves participants' affective states and their perception of the feedback providers. Platform designers could decide which activity to use based on how they prioritize these measures and compatibility of the activity with the existing workflow. Future work could explore the effectiveness of other coping activities and incorporate them into this design space.

Platform designers should implement an expressive writing activity if revision outcomes are the priority. For example, promoting revision outcomes may be most important in learning contexts where content creators may need to demonstrate depth of revision for course credit. It may also be important in professional contexts where showing depth of revision is a critical part of managing client relations. To implement expressive writing, platform designers could allow users to write private comments on each piece of feedback or write their reactions on the set of feedback holistically [154]. Writing comments on feedback could also have additional benefits, such as enabling reflection to promote interpretation of the content.

If platform designers value users' affective states or their perception of the feedback providers most, then they should consider implementing an intervention that enables a brief distraction from negative feedback. For example, affective states might be most important for platforms that cater to novice content creators. Prior work shows reduced affective states negatively impacts learning outcomes such as new skill development and demotivates future participation [155]. The perception of the feedback and its providers are also important on platforms that promote social interactions between members. Higher evaluation of the feedback providers helps to avoid conflicts among members and increase future engagement [52]. A short distraction improves users' perception of the feedback providers while allowing their affective states to restore to more neutral levels. More positive affective states could further protect the relationships between members by reducing aggressive behaviors on the platform. For distraction, platform designers could present links to view related content or perform non-feedback tasks in the community or to browse one's own project histories. Additionally, platform designers could disallow postings of revised content for a short duration to suggest that the user should first perform tasks unrelated to the content revision.

	Neutral	Negative
Structure	A stronger ending is in order. Perhaps the author can come up with more gripping and distinguished ending material.	The ending is terrible. Perhaps the author should at least come up with some less boring and plain ending material.
Style	The first error in my opinion is the use of personal pronouns. I was taught not to use "I" or "me" in essays because it makes the essay sound less professional.	One big obvious error here is the use of personal pronouns. The first thing I learned from Writing 101 is to avoid using "I" or "me" in essays. Only newest rookie uses personal pronouns.
Content	I pretty much agree with all of the points you have made. I would add the argument of the protection of civil liberties.	I don't really buy any of the points you have made. I would add the argument of the protection of civil liberties to make this essay closer to being convincing, if that's ever possible.

Table 4.1: Feedback set for the revision task. In total, there are three pieces of authentic feedback on structure, style and content. Each piece has two versions with neutral (left) and negative (right side) tone. Every participant receives one piece of feedback from each row.

	happy				enthusiastic			
	sum sq	**df**	**F**	**PR(>F)**	**sum sq**	**df**	**F**	**PR(>F)**
valence balance	217.02	3	35.00	<0.01	155.71	3	22.91	<0.01
coping activity	13.62	3	2.20	0.09	8.47	3	1.25	0.29
v:c	17.23	9	0.93	0.50	19.14	9	0.94	0.49
residual	958.93	464			1050.97	464		
	annoyed				**frustrated**			
valence balance	497.29	3	50.94	<0.01	310.61	3	34.93	<0.01
coping activity	0.89	3	0.09	0.96	6.84	3	0.77	0.51
v:c	25.61	9	0.87	0.55	28.28	9	1.06	0.39
residual	1509.80	464			1375.27	464		
	edit duration				**edit distance**			
valence balance	371.70	3	5.71	<0.01	8.88e5	3	5.80	<0.01
coping activity	391.10	3	6.01	<0.01	3.27e5	3	2.14	0.09
v:c	279.01	9	1.43	0.17	5.09e5	9	1.11	0.35
residual	1.01e4	464			2.36e7	464		
	positive				**useful**			
valence balance	598.58	3	118.76	<0.01	390.33	3	47.95	<0.01
coping activity	6.34	3	1.26	0.29	28.95	3	3.56	0.01
v:c	7.68	9	0.51	0.87	30.95	9	1.27	0.25
residual	779.53	464			1258.93	464		
	fair				**considerate**			
valence balance	462.88	3	62.70	<0.01	834.84	3	152.75	<0.01
coping activity	16.23	3	2.20	0.09	18.14	3	3.32	0.02
v:c	19.03	9	0.86	0.56	15.22	9	0.93	0.50
residual	1141.80	464			845.30	464		
	polite				**knowledgeable**			
valence balance	1022.71	3	195.23	<0.01	360.83	3	61.59	<0.01
coping activity	14.71	3	2.81	0.04	9.28	3	1.58	0.19
v:c	11.10	9	0.71	0.70	21.72	9	1.24	0.27
residual	810.23	464			906.13	464		
	expertise				**feedback accepted count**			
valence balance	240.58	3	38.64	<0.01	36.81	3	11.43	<0.01
coping activity	10.69	3	1.72	0.16	4.51	3	1.40	0.24
v:c	9.26	9	0.50	0.88	11.65	9	1.21	0.29
residual	963.07	464			498.23	464		
	original rating				**revised rating**			
valence balance	2.81	3	1.03	0.38	0.68	3	0.25	0.86
coping activity	4.75	3	1.74	0.16	3.87	3	1.42	0.24
v:c	6.39	9	0.78	0.64	6.04	9	0.74	0.67
residual	422.98	464			420.64	464		

Table 4.2: ANOVA results for all measures.

	happy	enthusiastic	annoyed	frustrated
0-	4.48^{a}	4.18^{a}	2.03^{a}	1.93^{a}
1-	3.85^{b}	3.54^{b}	3.29^{b}	2.87^{b}
2-	3.14^{c}	2.97^{c}	4.56^{c}	3.99^{c}
3-	2.72^{d}	2.69^{c}	4.41^{c}	3.70^{c}
control	3.40^{a}	3.27^{a}	3.63^{a}	3.24^{a}
self-affirmation	3.59^{ab}	3.36^{a}	3.58^{a}	2.92^{a}
expressive writing	3.40^{a}	3.19^{a}	3.55^{a}	3.13^{a}
distraction	3.81^{b}	3.55^{a}	3.52^{a}	3.18^{a}
LSD	0.36	0.38	0.45	0.44

Table 4.3: Mean affective state ratings across valence balance and coping activity conditions. The label "[0, 1, 2, 3]-" refers to the valence balance condition in which participants received [0, 1, 2, 3] pieces of negative feedback. Fisher's LSD (Least Significant Difference) [152] is used for post hoc test. Means with different superscripts are significantly different ($p<.05$). For example, 3.40^{a} is significantly different from 3.81^{b}. Nether 3.40^{a} nor 3.81^{b} is significantly different from $3.59^{a}b$.

	edit duration	edit distance
0-	6.11^{a}	260.6^{a}
1-	4.54^{b}	187.3^{b}
2-	4.25^{b}	181.2^{b}
3-	3.77^{b}	141.5^{b}
control	6.11^{a}	229.3^{a}
self-affirm.	4.29^{b}	166.7^{b}
exp. writing	4.61^{b}	205.4^{ab}
distraction	3.64^{b}	169.2^{b}
LSD	1.18	57.3

Table 4.4: Mean edit duration and distance across valence balance and coping activity conditions.

	pos.	useful	fair
0-	4.56^{a}	5.73^{a}	5.91^{a}
1-	3.30^{b}	4.83^{b}	4.84^{b}
2-	2.31^{c}	3.94^{c}	4.01^{c}
3-	1.58^{d}	3.34^{d}	3.28^{b}
control	2.98^{a}	4.73^{a}	4.60^{ab}
self affirmation	2.82^{a}	4.40^{ab}	4.48^{ab}
expressive writing	2.85^{a}	4.08^{b}	4.23^{a}
distraction	3.11^{a}	4.60^{a}	4.73^{b}
LSD	0.3	0.42	0.40

Table 4.5: Mean feedback perception ratings across valence balance and coping activity conditions.

	considerate	polite	knowledge.	expert
0-	5.74^{a}	5.83^{a}	5.38^{a}	4.38^{a}
1-	4.07^{b}	3.98^{b}	4.30^{b}	3.40^{b}
2-	2.95^{c}	2.73^{c}	3.63^{c}	2.97^{c}
3-	2.25^{d}	1.97^{d}	3.05^{d}	2.46^{d}
control	3.85^{ab}	3.76^{ab}	4.25^{a}	3.49^{a}
self-affirmation	3.63^{a}	3.46^{a}	3.91^{a}	3.18^{a}
expressive writing	3.52^{a}	3.44^{a}	4.01^{a}	3.14^{a}
distraction	4.02^{b}	3.83^{b}	4.20^{a}	3.40^{a}
LSD	0.34	0.34	0.35	0.37

Table 4.6: Mean feedback provider ratings across valence balance and coping activity conditions.

CHAPTER 5: VALENCE-BASED ORDER

The prior experiment tested self-directed activities to promote resilience to negative feedback. In this chapter, I explore the effects of a new technique that orders the presentation of a collection of feedback based on its sentiment.

In formal learning environments, instructors recommend the use of mitigating language, such as praise or affirmation, before or after negative feedback to improve its receptivity [86, 87, 156, 157, 158]. Similar techniques are less applicable in online environments, where the feedback is often composed by multiple independent providers and platform designers have limited control over the composition process. Prior work has tested various methods that improve the positive valence of feedback, including the use of rubrics [13] and positive examples [14]. While these methods may help, they do not eliminate occurrences of negative feedback. One common solution among platform designers is to remove the negative feedback [159, 160]. Despite its simplicity, this approach has several disadvantages. First, negative feedback may still contain constructive advice useful for learning and content improvement [161]. Second, removing feedback without consent may discourage the provider from participating further [162]. Third, feedback removal may be inapplicable in certain situations, such as when content creators have paid in advance for each piece of feedback.

In this chapter, we test a novel approach that presents a collection of feedback in an order from most-to-least positive valence. Specifically, our work examined whether positive feedback could be used to mitigate the influence of the negative feedback in the set. If effective, the technique could be automated in existing feedback platforms using sentiment analysis [163]. Prior work also shows that cues about the source (provider) of the feedback can affect perceptions of the feedback [43]. Our experiment additionally examined how the perceived source of the feedback (peers vs. experts vs. anonymous) [27, 89] affects content creators' reactions to the feedback, and how it mediates the effect of valence order.

We conducted an online experiment in which participants (n=270) wrote a children's story based on an illustration. Two days later they revised their stories based on a set of given feedback. The feedback set included two pieces of feedback with positive valence and one piece with negative valence. The feedback was presented with different valence orders (negative first, negative last, and negative between) and source identity cues based on the experiment condition. We measured participants' affective states, perceptions of the feedback and its source, revision extent, and story quality.

Our results showed that when the negative feedback was presented later (further down in the results), the content creator invested more effort in the revision task and rated the collec-

tion of feedback more favorably. Feedback source had no statistical effect on the measures. We also observed a gender effect. Female participants were more likely to accept feedback (reported applying it in the revised story) and the negative feedback causes a larger reduction in their affective states. Our work contributes deeper empirical understanding of how valence order and source identity can be used to improve feedback receptivity in online environments.

5.1 METHODOLOGY

Our experiment addresses three research questions:

- How does ordering a feedback set based on valence affect the extent and quality of the subsequent content revisions?
- How does valence order influence a content creator's affective state, and influence his or her perceptions of the feedback and its source?
- How does information about the source of the feedback providers affect these same measures, and mediate the effects of the feedback ordering based on valence?

Answers to these questions help content creators better utilize feedback received online (e.g., positive affective state is associated with increased creative thinking [84]). Also, the answers can help platform designers know how to more effectively present the feedback (e.g., how to order it and whether to display source cues).

5.1.1 Experiment Design

To answer the research questions, we conduct a 3x3 full factorial online experiment with two factors: Valence Order and Source. Each participant receives three pieces of feedback, including two with positive valence levels and one with negative valence level. There are three levels in Valence Order: negative first, negative between, and negative last. There are also three levels in Source: peer, expert, and anonymous. In the peer and expert conditions, the task instruction clearly states that the feedback comes from peer workers or domain experts. The feedback text also starts with the words "Peer Worker" or "Domain Expert" (see Figure 5.1). In the anonymous condition, participants receive no information about the feedback source. There is also no source identity cue in the feedback text.

Figure 5.1: Task interface for the revision phase. Feedback is provided in different orders and with different source cues based on experimental conditions. Participants have already read the feedback piece by piece before reaching this stage.

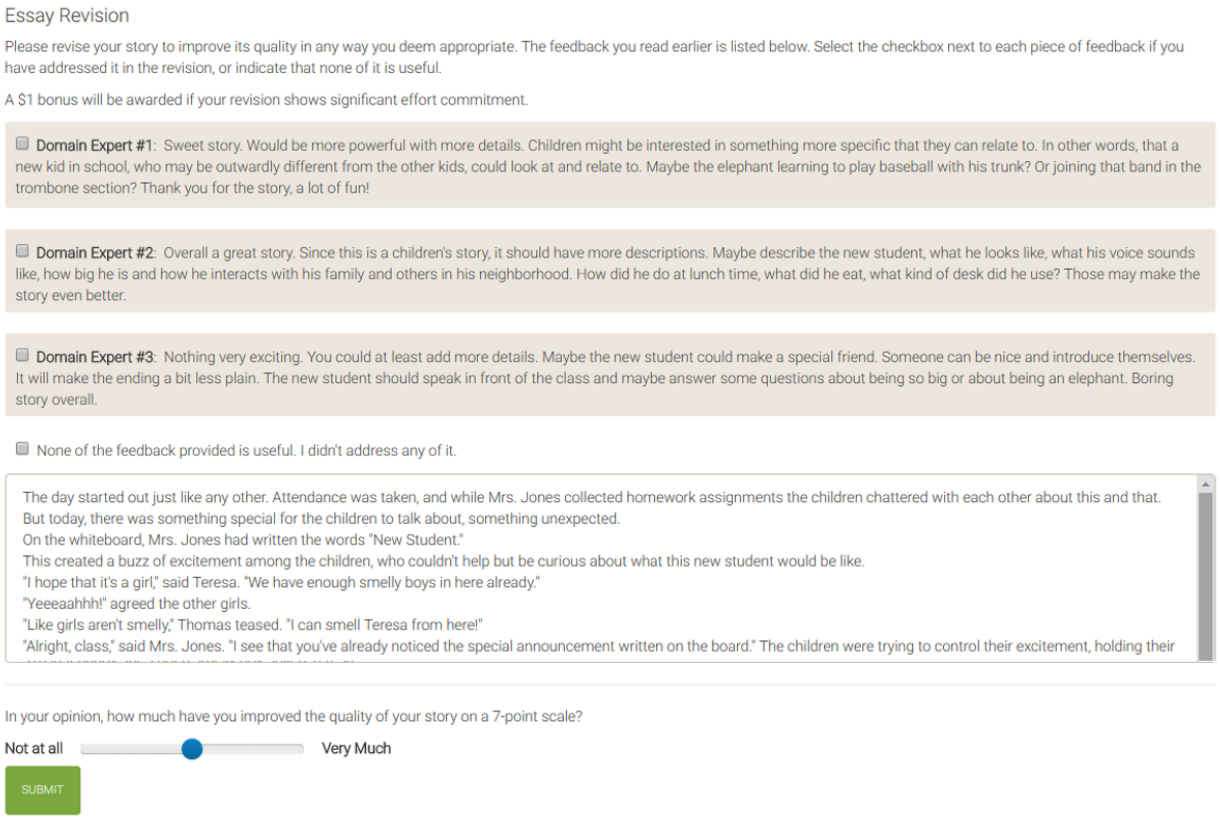


Figure 5.2: The user interface during the story writing phase.

Task Instruction

Please write a short story for children of 8-12 years old based on the illustration. Use your imagination to develop a creative plot. To help you write the story, you may consider answering the following questions:

- What is happening at the moment?
- How come the new student is an elephant?!
- What will happen shortly after?
- How does your story start and end?

Your story should be 200 words minimum and 2000 words maximum. Stories that are too vague or plagiarized from other sources will be rejected.

By writing a story of good quality, you may earn the opportunity to participate the second phase of the study, which offers the same amount of payment and potential bonus.

Stampy was a very smart elephant and they knew that at the local zoo

SUBMIT

5.1.2 Essay Task

The task includes two phases: story writing and revision. For the story writing phase, we ask participants to write a story for children of 8-12 years old based on a given illustration (Figure 5.2). A pilot study shows that most participants could finish the story in less than one hour. We intentionally allocate extra time for the task so the story quality would not be compromised due to time pressure. The participants have two hours to write a story within a 200-2000 word limit. We choose story writing as our experimental task for three reasons; (i) it is a topic that should be familiar to a general audience; (ii) it requires creative thought; and (iii) it only requires text entry, making it suitable to perform online.

The illustration facilitates the task in two ways. First, it provides scaffolding in the open-ended writing process by outlining the story's main characters and scenario. Second, it allows the research team to select general feedback that applies to most stories by narrowing the scope of the possible story themes [164]. An external expert in story writing selects the illustration based on task appropriateness. Participants receive $3 for the writing phase.

In the story revision phase, participants revise stories based on a set of feedback. After reviewing their stories at the start of the phase, participants receive three pieces of feedback. The task interface presents one piece of feedback at a time. Participants select a button to reveal the next piece of feedback. After feedback delivery, we ask participants to complete a survey about their affective states and perceptions of the feedback set and its perceived source. Then participants revise their original story to improve its quality. Participants receive an additional $2 for the revision phase. To discourage satisficing, we offer a $1 bonus if they demonstrate significant effort during the revision phase. In total, top 30.9% of all participants ranked by edited character count receive the bonus.

5.1.3 Feedback Pool

The feedback assigned to participants came from a feedback pool consisting of six pairs of positive and negative feedback (Table 5.1). Each pair was adapted from one piece of authentic feedback collected online. We used five stories from a pilot study to collect a large set of authentic feedback on the story plots. From the set, we selected six pieces of feedback that gave revision advice on story content. The feedback type was decided based on prior work about effective feedback [165].

To ensure each piece of feedback suggested a similar degree of revision, we recruited 30 online judges from Amazon Mechanical Turk to rate the actionability levels. Each judge reviewed 7 pieces of feedback including the 6 pieces of authentic feedback and one duplicate

for quality control. For each piece of feedback, the judge rated the extent of revision needed if the feedback was accepted on a 7-point scale from 1 (No Revision Needed) to 7 (Major Revision Needed). We discarded the ratings from judges who rated the duplicate piece of feedback noticeably different (larger than two units) from its counterpart. The final average actionability rating across the feedback set was 4.12 (SE=0.26), and there was no significant difference between the ratings of any two pieces of feedback. The valence levels of the feedback pool were also adjusted and validated in the same manner. In the end, all adjusted positive feedback have similar positive valence levels (μ=5.38, SE=0.26), and the negative feedback have similar negative valence levels (μ=2.68, SE=0.20).

Positive Valence Version	**Negative Valence Version**
I really loved the story. I would change a couple of things. The story would be more interesting if more details were given. For example, what happened to his parents? In what ways does he feel different from the children? Does he miss living like an elephant? I teach young children and I would read this story to my class.	A pretty boring story. I would change a couple of things. The story would be more interesting if more details were given. For example, what happened to his parents? In what ways does he feel different from the children? Does he miss living like an elephant? I teach young children and I may not read this story to my class.
Great story! I think the new student also needs to introduce himself to the class so they can learn more about him. He can tell them where he is from, about how it is different from his new home area, what he likes to do, etc, so they can get to know him. The teacher can also ask the classmates to speak up if there is anything they like that the new student likes. That may make both the new student and the classmates more comfortable and willing to accept each other.	Quite a boring story. Some more details may make the story less plain. I think the new student also needs to introduce himself to the class so they can learn more about him. He can tell them where he is from, about how it is different from his new home area, what he likes to do, etc, so they can get to know him. The teacher can also ask the classmates to speak up if there is anything they like that the new student likes. That may make both the new student and the classmates more comfortable and willing to accept each other.

Table 5.1: The feedback pool from which the research team assigned three pieces of feedback to each initial story. At most one piece of feedback was assigned from each feedback pair (each row). The left and right columns show the positive and negative valence versions of the feedback, respectively.

Overall a great story. Since this is a children's story, it should have more descriptions. Maybe describe the new student, what he looks like, what his voice sounds like, how big he is and how he interacts with his family and others in his neighborhood. How did he do at lunch time, what did he eat, what kind of desk did he use? Those may make the story even better.	Overall a pretty boring story. Since this is a children's story, it should at least have more descriptions. Maybe describe the new student, what he looks like, what his voice sounds like, how big he is and how he interacts with his family and others in his neighborhood. How did he do at lunch time, what did he eat, what kind of desk did he use? Those may make the story less boring.
Great story. You should add more details to make it even better. I would like for the new student to make a special friend. Maybe someone can be nice and introduce themselves. It will be an even happier ending to the story. The new student should speak in front of the class and maybe answer some questions about being so big or about being an elephant. Good job overall!	Nothing very exciting. You could at least add more details. Maybe the new student could make a special friend. Someone can be nice and introduce themselves. It will make the ending a bit less plain. The new student should speak in front of the class and maybe answer some questions about being so big or about being an elephant. Boring story overall.
Sweet story. Would be more powerful with more details. Children might be interested in something more specific that they can relate to. In other words, that a new kid in school, who may be outwardly different from the other kids, could look at and relate to. Maybe the elephant learning to play baseball with his trunk? Or joining that band in the trombone section? Thank you for the story, a lot of fun!	Boring story. Would be less plain with more details. Children might be interested in something more specific that they can relate to. In other words, that a new kid in school, who may be outwardly different from the other kids, could look at and relate to. Maybe the elephant learning to play baseball with his trunk? Or joining that band in the trombone section? I didn't really enjoy reading the story.

Table 5.1 cont.

My daughter may love this story. I would like to read a bit more about the elephant's first day in the classroom - how he sat down, how the other children reacted, how he participated in the classroom work, and how the teacher treated him. I think that those details might add some more color to the story and perhaps even a bit more tension.	I probably wouldn't read this story to my daughter, just too boring. Maybe you can talk a bit more about the elephant's first day in the classroom - how he sat down, how the other children reacted, how he participated in the classroom work, and how the teacher treated him. I think that those details might add some more color to the story and perhaps even a bit more tension.

Table 5.1 cont.

5.1.4 Participants

In total, 270 participants complete the experiment. All participants are recruited from Amazon Mechanical Turk and located in the U.S. All participants have finished more than 500 HITs and a minimum 95% approval rate. Among the participants, 40% are males, 60% females; 98.1% report English as their first language; 86.3% have read stories to children; 39.3% are parents of children younger than 15 years old. Regarding the age distribution, 9.6% were 18-24 years old, 42.2% 25-34 years old, 27.4% 35-44 years old, and 20.7% 45 years or older. Regarding the education level distribution, 40.8% report high school or lower as their highest academic degree earned, 44.8% undergraduate degree, and 14.4% graduate degree. Since most workers on AMT earn at or below minimum wage [166], we assume that the participants were novice content creators rather than professional writers.

5.1.5 Procedure

Participants read an IRB consent form and fill out a demographic survey at the beginning of the experiment. The task instruction also informs the participants about the following revision phase. Then we give participants two hours to compose their initial stories. After all stories are collected, the research team selects three feedback pairs from a 6-pair feedback pool for each story based on their appropriateness for the plot (Table 5.1). Each pair includes one piece of positive and one piece of negative feedback, and both pieces are derived from the same piece of authentic feedback. For the three feedback pairs selected by the research team, a Python script randomly selects one piece of feedback from each pair, and two piece

of positive feedback and one piece of negative feedback in total. For 22.3% of the stories, there are fewer than three feedback pairs applicable to the plot and we thus exclude these stories from the revision phase. While assigning feedback for all stories, we monitor how many times each feedback pair has been selected and adjust to ensure an even allocation of the 6 pairs. Two days after participants finish the writing phase, we launch the revision phase and notify the participants via email. 74.6% of all qualified participants complete the revision phase. The task presents the feedback in different valence orders and with different source cues based on experiment conditions (Figure 5.1). The participants have two hours to finish the revision phase.

5.1.6 Measurements

We collected three sets of measurements:

- Affective states: how distressed / upset / enthusiastic / inspired / excited / happy they felt after receiving the feedback.
- Perceptions of the feedback and its source: how useful / positive / fair they perceived the feedback to be, how knowledgeable / polite the feedback sources to be, and how good they perceive their writing skill to be after receiving the feedback.
- Revision: how much time participants spent writing the initial story, reading the feedback, and revising the story; how much the story changed during the revision in terms of self- and expert-rated quality improvement, and edit distance between the initial and revised stories.

We collected the first two sets of measurements from the survey, which included 13 statements regarding participants affective states (6 items adapted from PANAS [150]), their perceptions of the feedback (3 items) and its source (2 items), and confidence in writing skills (2 items). Metrics related to revision extents were derived from participants' action logs. For the quality improvement rating, two experts in English writing each rated all 270 stories. The rating interface presented both the initial and revised versions side by side, and the experts rated how much the revision had improved the quality of the story on a 7-point scale (-3: the original has much higher quality; +3: the revised has much higher quality; 0: no noticeable quality difference). For 91.7% of all stories, the rating difference between the experts was smaller or equal to one unit on the scale. We averaged the two ratings as the final quality improvement. The average quality improvement was 1.37 (SE=0.17) across all conditions.

5.2 RESULTS

In the following subsections, we report the most interesting patterns in our data. Figure 5.3 summarizes the results.

5.2.1 Participants are most motivated when reading negative feedback last.

An ANOVA shows that Valence Order has a main effect on the ratings of enthusiasm ($F[2, 267]=3.96$, $p=.02$), excitement ($F[2, 267]=3.51$, $p=.03$), and happiness ($F[2, 267]=3.61$, $p=.03$). See Figure 5.4. Participants report significant higher ratings in the negative last condition (enthusiasm: $\mu=4.96$, $SE=0.17$; excitement: $\mu=4.73$, $SE=0.18$; happiness: $\mu=5.01$, $SE=0.16$) than in the negative first condition (enthusiasm: $\mu=4.30$, $SE=0.16$; excitement: $\mu=4.10$, $SE=0.17$; happiness: $\mu=4.46$, $SE=0.16$; $p<.05$). Simply presenting the same feedback in different orders increases participants' enthusiasm by 11.0%, excitement by 10.5%, and happiness by 9.2% on a 7-point scale. In general, the later in the order participants read the negative feedback, the more enthusiastic, more excited, and happier they are. In the negative last condition, participants may view the first two pieces of positive feedback as an affirmation of the quality of their stories and become more resilient to the influence of the negative feedback. There is no significant effect of Source on participants' affective states.

5.2.2 Participants report most favorable perception of the feedback when reading the negative feedback last.

Participants in negative last condition rate the feedback set marginally fairer ($\mu=5.70$, $SE=0.14$) than in the negative first condition ($\mu=5.33$, $SE=0.14$; $p=.06$). The feedback positivity and usefulness ratings also show the same trends, but does not reach statistical significance. On average, participants rate the feedback fairer by 6.2%, more positive by 5.3%, and more useful by 5.2% on a 7-point scale in the negative last condition in comparison with the negative first condition. Source does not have a statistical effect on participants' perception of the feedback measured in the experiment.

5.2.3 Anonymity tends to improve affective states and perceptions of feedback and its source.

Despite the lack of statistical significance, we observe interesting trends consistent with prior work regarding feedback source anonymity [27]. Our results show source anonymity

Figure 5.3: Measurements collected during the experiment. For all charts, the left / center / right groups of bars represent results from negative feedback presented first / between / last conditions. We also color-coded all bars according to source conditions: expert, anonymous (darkest), and peer (lightest). The vertical axes cover both the minimum and maximum (if applicable) range of the measurements.

tends to improve participants' affective states and lead to more favorable perceptions of the feedback set and its source. In the anonymous condition, participants report being 4.8% more enthusiastic, 5.7% more inspired, 4.6% more excited, 2.9% happier, 4.4% less distressed, and 2.6% less upset than in the peer and expert conditions. Participants also rate the feedback 4.1% more useful, 2.2% more positive, 3.4% fairer, and the feedback source 2.7% more knowledgeable, 1.4% politer on a 7-point scale in the anonymous condition.

Contrary to prior work [89], our results show there is no statistical difference in task performance between the expert and peer cue conditions. Participants in the expert condition spend more time reviewing the feedback (μ=94.3 sec, SE=12.2) than in the peer condition (μ=68.8 sec, SE=12.2), but they perceive the source in the peer condition (μ=4.90, SE=0.15) to be nearly as knowledgeable as in the expert condition (μ=5.02, SE=0.15). There is also no statistical difference in the edit distance or quality improvement.

One potential reason may be participants' familiarity with the writing task. 86.3% of the participants report having told stories to young children. Participants with storytelling experience self-report significantly higher quality improvement (μ=5.38, SE=0.08) in comparison with the participants without (μ=4.89, SE=0.23; $t(268)=2.17$, $p=.031$). Participants familiar with the task were more confident in their performance and their writing skill (participants w/ exp.: μ=3.78, SE=0.10; w/o: μ=3.41, SE=0.26). Prior work shows people with higher self-efficacy are less receptive to feedback [167]. The participants' familiarity with the task domain may therefore have affected how they perceived the source cues.

5.2.4 Improved affective states and feedback perception lead to more revision.

Neither Valance Order nor Source has a significant effect on the edit distance (μ=1327.56, SD=1025.82), feedback accepted count (μ=2.12, SD=0.06), and expert-rated quality improvement (μ=1.37, SD=0.86). Edit distance is significantly correlated with the feedback accepted count (Pearson's $r=.41$; $p<.01$) and quality improvement ($r=.73$; $p<.01$). The more participants edited their essays, the higher the quality improvement ratings are (Table 5.2). We also observe significant but weak correlations between the edit distance and ratings of enthusiasm ($r=.14$; $p<.05$), excitement ($r=.15$; $p<.05$), and inspiration ($r=.17$; $p<.01$). More motivated participants tend to revise their work more; therefore methods to improve motivation such as ordering feedback by valence as done in this study, providing immediate positive feedback [84], or wrapping feedback with positive language [27] as done in prior work can foster revision. Interestingly, participants' distress level also has a positive correlation with the edit distance ($r=.13$, $p<.05$). This correlation may be caused by the acceptance of negative feedback, which increases the edit distance and distress level at the same time.

Participants are more likely to accept feedback that leaves more favorable impressions. Edit distance has significant correlations with how useful (r=.18, p<.01), positive (r=.12, p<.05), and fair (r=.21, p<.01) participants perceive the feedback set to be.

	edit distance	**accepted count**	**expert quality improvement**
happy	-0.03	0.05	-0.07
upset	0.1	-0.02	0.05
distress	0.13	-0.02	0.1
excited	0.15	0.13	0.06
enthusiasm	0.14	0.18	0.06
inspired	0.17	0.18	0.07
feedback positive	0.12	0.21	0.11
feedback useful	0.18	0.19	0.17
feedback fair	0.21	0.24	0.27

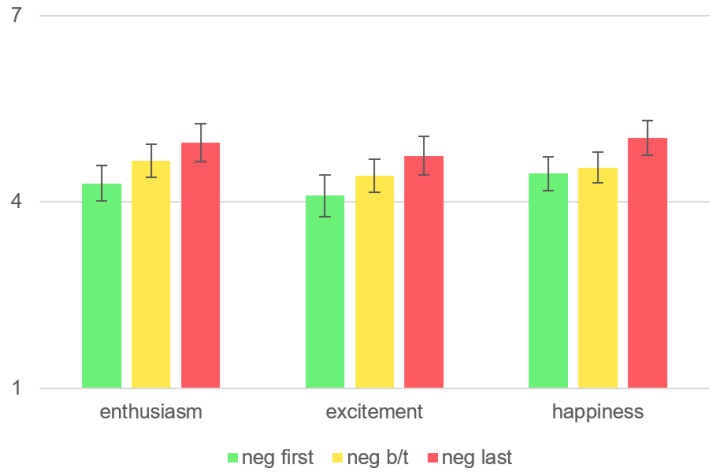
Table 5.2: Correlation table between revision extent metrics and participants' affective states and feedback perception. More favorable perception of the feedback set and more positive affective states correlate with a greater degree of revision.

5.2.5 Female participants were more receptive to feedback

Table 5.3 shows the gender difference in our results. Prior work finds women are more influenced by verbal evaluative feedback than men [168]. In our experiment, female participants accepted significantly more pieces of feedback (μ=2.23, SE=0.07) and spent marginally longer time reading feedback (μ=1.62 min, SE=0.17) than male participants (μ=1.94, SE=0.09, t(268)=-2.53, p=.012; μ=1.16 min, SE=0.15, t(268)=-1.92, p=.056). Female participants also edited their stories more (μ=1495.35, SE=80.53), spent more time editing (μ=28.65, SE=1.93), and reported higher self-rated quality improvement (μ=5.49, SE=0.09) than male participants (μ=1075.87, SE=93.78, t(268)=-3.35, p<.001; μ=21.77, SE=2.32, t(268)=-2.27, p=.024; μ=5.03, SE=0.14, t(268)=-2.92, p=.004 respectively).

On the other hand, negative feedback had a stronger influence on female participants' affective states. Female participants reported feeling more distressed (μ=2.72, SE=0.12), more upset (μ=2.58, SE=0.12), and marginally less happy (μ=4.54, SE=0.14) compared to male participants (μ=2.17, SE=0.14, t(268)=-2.85, p=.005; μ=2.05, SE=0.13, t(268)=-2.89, p=.004; μ=4.88, SE=0.15, t(268)=1.86, p=.064). In sum, female participants were more likely to accept and be influenced by the feedback.

Figure 5.4: This chart shows the ratings for enthusiasm, excitement, and happiness; clustered by feedback ordering. Presenting negative feedback last resulted in higher ratings of participants' affective states.



	Female	**Male**
accepted count *	2.23 (.07)	1.94 (.09)
feedback reading time	1.62 (.17) min	1.16 (.15) min
edited char count **	1495.35 (194.7)	1075.87 (147.7)
story editing time *	28.65 (1.93) min	21.77 (2.32) min
distressed **	2.72 (.13)	2.17 (.14)
upset **	2.58 (.12)	2.05 (.13)
happy	4.54 (.11)	4.88 (.15)

Table 5.3: Gender comparison between feedback receptivity and affective states. Standard errors of the means are included in parenthesis. For each row, '*' indicates significance level of $p<.05$, and '**' indicates $p<.01$.

5.2.6 Manipulation check

The design of the experiment assumes that participants would read the feedback top to bottom — in the order that it is presented for each experimental condition. It is possible that content creators selectively read the feedback in orders of their preference, regardless of how the results are ordered on the screen. To test this possibility, we conduct a post hoc experiment on AMT where participants review three pieces of feedback and self-report the order they follow. Participants first review the same design as the one used for the original study in a task interface identical to the one used in the main experiment. Then they move on to a second screen that presents three pieces of feedback (two positive ones and one negative). Lastly, they answer a single-choice question to report the order in which they read the feedback. See Figure 5.5. Out of the 30 participants, 29 report reviewing the feedback from top to bottom. The other one selects "other" orders without giving a specific explanation. These results indicate participants in the main study should follow the same pattern and review the feedback in the provided order.

Another question is about the feedback set used in the study. Here we have six feedback pairs with positive and negative valence variations. Although we have fine-tuned the language so all six pieces of negative feedback have similar levels of negative valence, the experiment did not control for feedback that was directed at the content vs. directed at the creator of that content. To examine whether there is a difference between feedback directed at the content vs the creator, we modified each piece of feedback in the pool to be directed at the content or its creator. Thirty participants recruited from AMT review the feedback and rate the perceived valence on a seven-point scale (Figure 5.6). Our results show there is no significant difference between the two (person: mean=2.67, sd=1.69; design: mean=2.83, sd=1.15). In other words, the uncontrolled balance of content and creator targeted feedback should not affect our main results significantly. Meanwhile, the positive feedback indeed received a significant higher rating (mean=5.57, sd=0.94) that the negative ones.

5.3 DISCUSSION

Our results show presenting negative feedback last improves participants' affective states and perception of the feedback set. A post hoc experiment confirmed that participants likely read the feedback in the same order as it was presented. Cues of the feedback source have no significant effect on participants' affective states, and no effect on perceptions of the feedback set and its providers. There is no interaction between valence order and feedback source.

In our experiment, participants receive three pieces of feedback for the initial story. This

makes the experiment tractable and gives the necessary control, but future work should test whether our results generalize to different sizes and valence balances of a feedback set. For larger feedback sets, platform designers could choose to select representative pieces of the feedback to summarize the larger set. Some online platforms, such as Amazon.com, have already adopted this method by showing the highest rated positive and negative reviews as a summary. This presentation mechanism allows users to quickly grasp the key insights without spending significant time consuming all reviews. Online design communities may explore similar techniques based on the valence level and the popularity of the feedback.

Similar techniques could also be used in creativity support tools for writing. When presenting comments collected from external reviewers, writing support tools could offer positive valence comments first and negative ones last. In the case where there is a large quantity of comments or when it is difficult to re-order the feedback (e.g., for inline comments), the tool could show only the positive comments as the default and users could access the additional feedback through interaction. On the other hand, tools could prompt feedback providers to write separately about the positive aspects of the work, and display a summary of these responses first. Future work could also test data-driven approaches that automate positive valence feedback. The system could compare content creators' performance, in terms of grammatical error rate or estimated vocabulary size, against their own prior writing or their peers, and report the positive results.

In our experiment, we achieve different levels of valence by adjusting the language of the feedback. Our results may also generalize to other visual indicators of valence. Platform designers in creative domains could take inspiration from other online review services. Some online work platforms deliver feedback along with valence indicators such as upvote/downvote in performance review or job approval/rejection scenarios. These indicators make it straightforward for platform designers to order feedback by valence. Another common form of a valence indicator is a numeric rating such as star ratings or scores on review sites. Fine-grained ratings make it easy to compare the valence levels among feedback. Platform designers could implement these valence indicators in online feedback collection services and facilitate the valence ordering process.

Feedback valence order has the same influence across the source conditions in our study. The valence order may therefore have similar effects on platforms where feedback providers have different social identities and expertise. The participants in our study are mainly novice content creators. Experienced content creators may react differently to the manipulations. Prior work shows experts seek negative feedback more actively than novices [169]. Negative feedback may therefore have a weaker impact on experienced content creators. Future work should test whether experience level of content creators interacts with valence order.

Figure 5.5: Screenshots of the manipulation check task: feedback presentation page (top); survey (bottom).

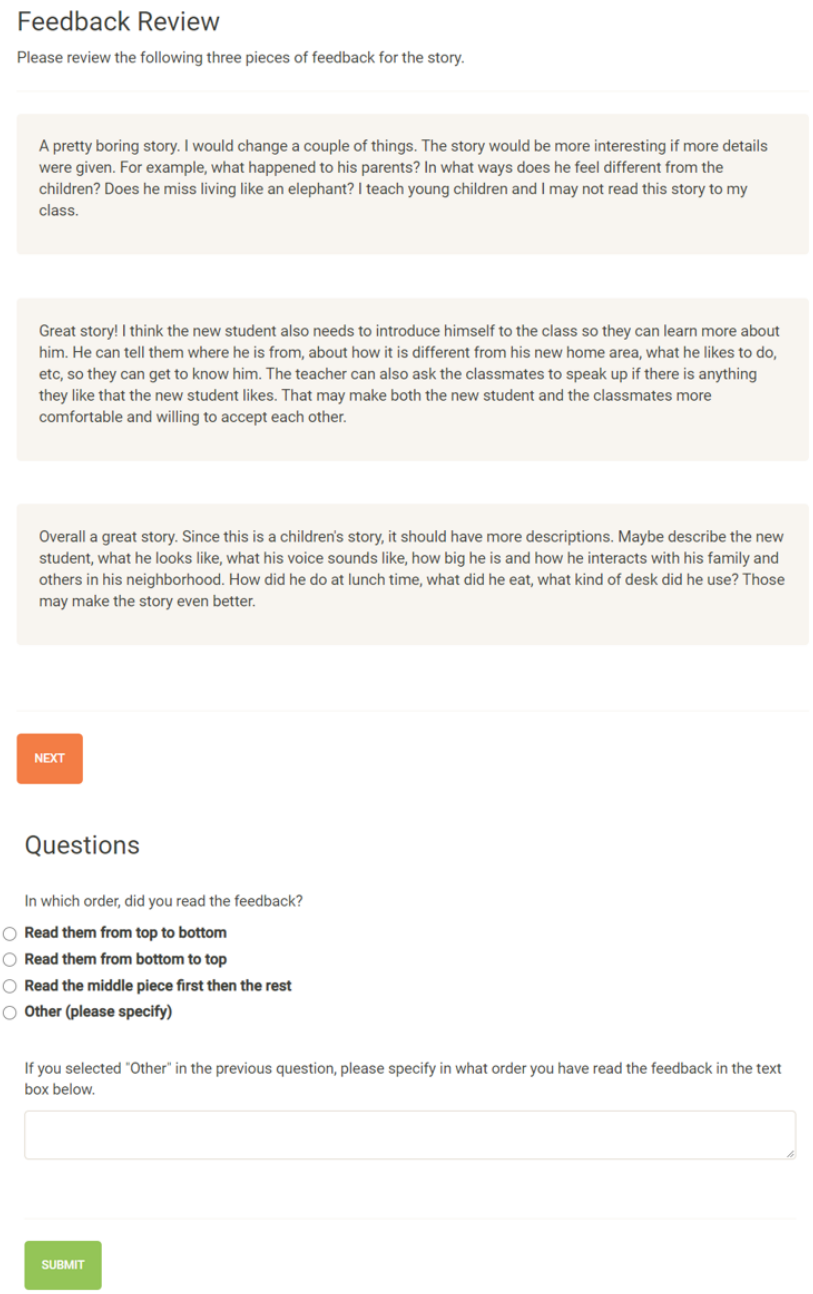
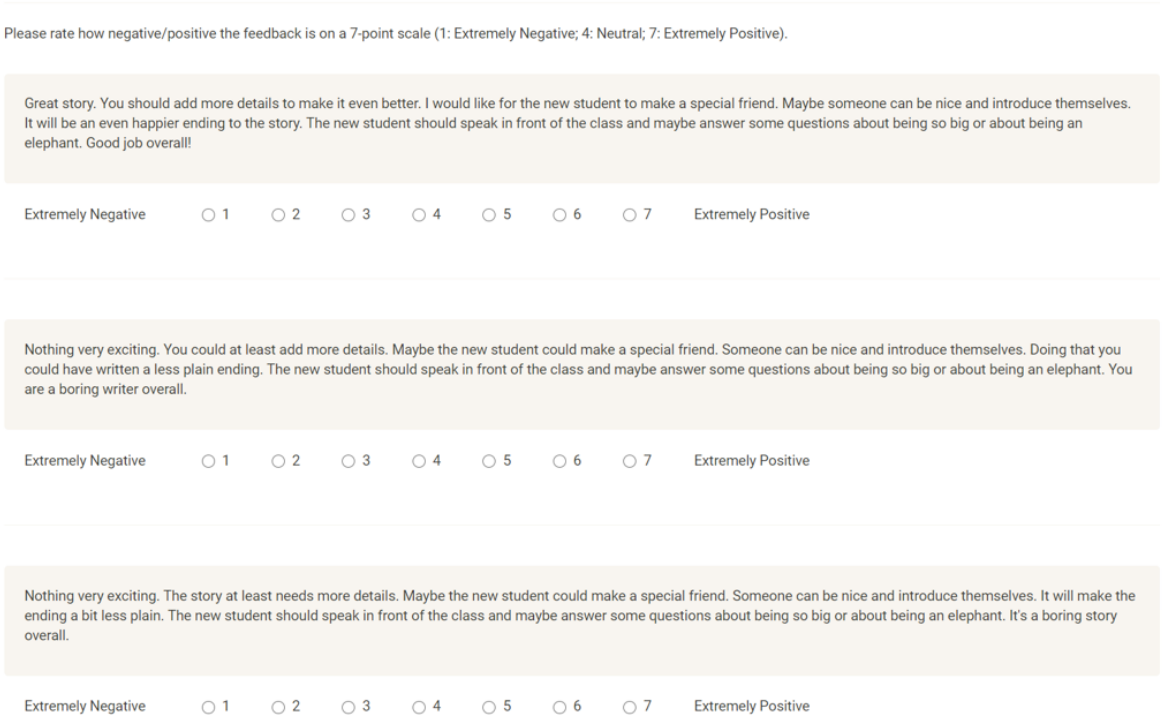


Figure 5.6: Screenshot of the manipulation check task. On this page, participants rate the perceived valence level of the feedback targeting at the work or the content creator respectively.



CHAPTER 6: EMPATHY AROUSAL & INGROUP FRAMING

The prior chapters focused on techniques directed at the content creator: one to promote resilience to negative feedback and the other to buffer the effects of negative feedback by presenting it last. In this chapter, we explore a technique aimed at the feedback provider – to increase their motivation to write constructive feedback. Even the most effective method cannot make content generators ignore the negative tone in the feedback completely while accepting the constructive advice. Therefore, for the last part of my thesis, I explore ways to discourage feedback providers from writing negative feedback in the first place. Specifically, we are interested in studying whether empathy arousal via narratives and ingroup framing could achieve such goals. Empathy is a vicarious response to another's emotional states. High level of empathy towards content creators may discourage the generation of negative feedback. Ingroup framing has close relationship with empathy. Perceiving others as ingroup members could increase the likelihood of empathy arousal. In this chapter, we hypothesize that platform designers should prioritize the relationship between the feedback provider and recipients. The combination of ingroup framing and empathy arousal mechanisms may induce higher quality feedback and written with a more constructive tone.

In this study, we study interventions that operationalize empathy to improve feedback quality and discourage negative feedback. Providing constructive feedback is a type of prosocial behavior. One empirically validated way to promote prosocial behavior is by eliciting empathy [131, 170, 171]. In our experiment, we examined two empathy-based approaches: narrative empathy and ingroup framing, and whether they could improve feedback quality and mitigate negative feedback. In the context of feedback exchange, empathy may encourage effort and thus higher quality feedback. It also reduces the likelihood of aggressive behaviors and promotes intervening action when negative feedback occurs. While there are many ways to arouse empathy, in our experiment, we chose narratives as they are applicable for most online platforms. Reading narratives from content creators encourages perspective-taking and improves the effectiveness of feedback [172]. Prior work indicates that empathy has two core aspects, affective empathy and cognitive empathy [173]. In our work, we examined the effectiveness of both aspects, testing narratives about experience receiving negative feedback (affective empathy) and design process of the target work (cognitive empathy).

We also examined how ingroup framing interacts with the empathy narrative interventions. Ingroup framing is one of the factors that affect empathy and there is extensive literature studying the interaction between these two [38, 39, 109, 110, 111, 112, 113, 114]. Prior work shows perceiving others as ingroup members promotes prosocial behaviors [174]. In addi-

tion, people are more likely to have empathetic feelings towards ingroup members. Following prior work, we implemented ingroup framing by establishing interdependency between the feedback provider and the designer and assigning a label to the pair [115]. Then we examined how ingroup framing interacted with narrative empathy interventions and influenced feedback composition.

We conducted a 3x2 (n=205) full factorial experiment with two factors: narrative empathy and ingroup framing. Narrative empathy has three levels: design process, negative experience, and control article. Ingroup framing has two levels: ingroup and control framing. In the experiment, participants reviewed a poster design, then read a passage with different content based on the narrative empathy condition, and later wrote feedback for the poster. Participants performed the task under a group framing or independently based on the ingroup framing condition. Afterward, we recruited domain experts to evaluate the quality of the collected feedback and used feedback length as a heuristic for effort. We also measured changes in participants' attitudes towards negative feedback in pre and post-study surveys.

Our results showed that both the design process and ingroup framing interventions significantly increased the effort invested in feedback composition by 40% and the final feedback quality by 30%. The negative experience condition had similar effects, increasing both measures by 20%. Also, the pre and post-study surveys showed participants experiencing the negative experience narrative condition reported a significantly more disapproving view towards negative feedback and significantly more likely to accept responsibility to intervene if negative feedback occurred, which could reduce the snowballing of negative feedback.

Our work makes three contributions to the CHI community: i) empirical evidence that empathy arousal and team dependency improve feedback exchange; ii) a deeper understanding of the underlying theories of narrative empathy, ingroup framing, and their interactions; iii) practical guidelines regarding how platform designers could use these interventions to help users to receive higher quality feedback while mitigating the prevalence of harsh criticism.

We focused on answering two research questions:

- R1: How do narratives such as reading about the designer's experience receiving negative feedback or the design process of the project influence feedback composition and attitudes towards negative feedback?
- R2: How does ingroup framing influence participants and interact with the effects of empathy-arousal narratives for the same measures?

We answered these two research questions through an online experiment.

Figure 6.1: A screengrab of the experimental task. Participants reviewed the poster and background information for the poster provided by the designer (e.g., the people shown were the honoree and chairs of the event). They then read a bonus opportunity statement, which was phrased differently based on the group framing conditions. Participants then wrote feedback targeting the perceived strengths and weaknesses of the poster and suggestions for improvement.

Your teammate created the poster for the Boys & Girls Clubs of Philadelphia to promote their annual spring fundraiser, the Philly Showcase of Wine, Cheese & Beer. Attendees have access to hundreds of fine wines, cheeses, beers and local food vendors. The four people in the center of the photo were the honoree and chairs of the event. The poster was intended to be displayed on walls throughout the city including public transportation.

> **Bonus**: To reward high performance, we offer a bonus opportunity in this HIT. We have grouped the designer of the poster and you as Team Orange and you two will collaborate on this task. After you submit the HIT, the the designer of the poster will review your feedback and try to revise the design accordingly. A domain expert will rate the success of your collaboration by evaluating the improvement in the design based on the feedback that you provide. The teams who score in the top 10 will share a cash bonus of $8, $4 per team member.

Please write your feedback in the text-box below. The feedback should incorporate the strengths and weaknesses of the poster, and provide specific and actionable suggestions for improvement.

6.1 METHODOLOGY

We conducted a 3x2 full factorial experiment with two factors: narrative empathy and ingroup framing. Narrative empathy had three conditions: negative experience, design process, and control article. Ingroup framing had two conditions: ingroup framing and control framing.

6.1.1 Experimental Task

In the online experiment, we asked participants to review a poster design and then provide feedback to help the designer to improve it (see Figure 6.1). We collected the design from a designer (Asian female with 2.5 years of experience at the time of the experiment) recruited from UpWork, a popular freelancing platform. Participants reviewed the design together with its background information, including its purpose, target audience, explanations of design elements, and where it would be displayed. Later, we instructed participants to write about the strengths and weaknesses of the design and provide actionable and specific advice to help the designer improve the design. We also promised a bonus for high-quality feedback to incentivize participants.

6.1.2 Narrative Empathy Factor

Narrative empathy had three conditions: negative experience, design process, and control article. For all three conditions, participants read a 300-word passage with content based on experimental conditions (see Table 6.1: negative experience, Table 6.2: design process, Table 6.3: control article. For negative experience, the passage described a prior episode of the designer receiving negative feedback on a creative project. From a first-person perspective, the designer recalled who commissioned the design, how s/he received the negative feedback, and how he s/felt at that moment. We recruited a designer (not the one who provided the design for the study) from UpWork to compose the passage based on personal experiences.

For the design process condition, the passage described the purpose of the design, design decisions made in the process, the reason to seek feedback, and how the designer felt about feedback collection. We asked the designer of the poster to provide the narrative. For both narratives, we edited the text for brevity and revised passive sentences into active forms, as prior work showed this style is more effective for arousing empathy [100, 133].

For the control article, the passage described a technological concept, a topic that was orthogonal to the task, from a third-person perspective. We selected the passage from a news

One day, I was told to create a design for an upcoming event hosted by a local client. The client preferred a minimalistic style, meaning it'd be a simple yet visually-appealing design. I quickly cranked out a couple mock-ups as this was a style that I had created similar designs before.

Later I visited the client to present my designs. I knew something was off when I saw their smile turn into a frown. Before I could ask for suggestions, I was immediately interrupted. "This is not at all what I'm looking for," they said. I've dealt with criticism in the past, but I was not prepared for their rude words. "It's pathetic and weak-looking." I turned my gaze from my designs to their face to see if they would laugh and say "just kidding," but their tone became more offensive as they began to harshly analyze my designs further. "Have you not learned anything from your time working with us? A rookie with no experience could have made a better mock-up than what I'm looking at." I was taken aback. Not quite yelling, but still louder than necessary, they continued in great detail while the others watched. "It's juvenile and low-quality. If you can't handle these simple projects, then maybe I need to find someone else who can."

These condescending words were coming from someone that I respected. The fact that it came from someone close to me who generally supported me and my work made the situation worse. I was visibly hurt, and embarrassed by the way that I was being addressed. It was unexpected and uncalled for. It's one thing to criticize my work, but to go after my skills and abilities and imply that I'm not good enough is extreme.

Table 6.1: Negative experience narrative: The passage described a prior episode of the designer receiving negative feedback on a creative project.

site and revised it to match the length of the other two narratives. For all three passages, we rephrased some sentences so they had the same level of readability (negative experience: 6.18; design process: 6.30; control article: 7.00 evaluated by Automated Readability Index [175]).

6.1.3 Ingroup framing factor

Group framing had two conditions: ingroup framing and control framing. Participants read different instructions throughout the task based on the conditions (see Table 6.4). For the ingroup framing condition, we informed the participants that we had grouped them and the poster designer as a team. Following prior work, the team was assigned an arbitrary and neutral label, namely "Team Orange". We also created interdependency between the participant and the designer by promising a team-based bonus. We informed the participants that a domain expert would rate the usefulness of their feedback by evaluating how much the designer improved the design using their feedback. The top 10 teams with the highest rating would share a bonus of 8,4 for each person. For the control framing condition, we instructed participants to complete the task independently. A domain expert will rate how useful their feedback was by evaluating its potential to help the designer improve the design. The participants writing the feedback rated in the top 10 would receive a bonus of $4. Throughout the experiment, we referred to the provider of the design as "your teammate" or "the designer" respectively based on the group framing condition.

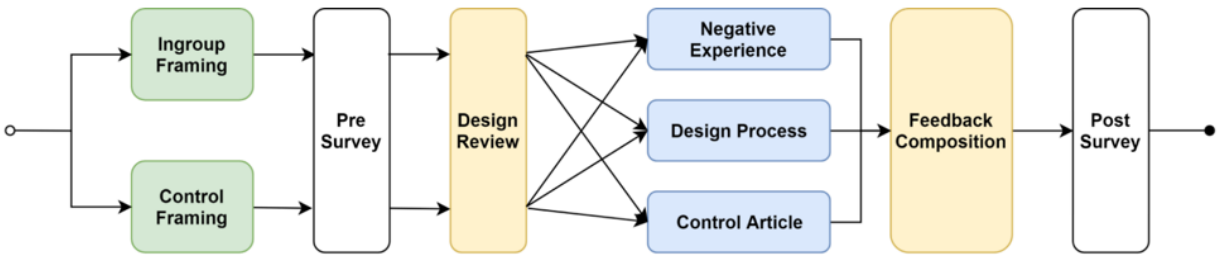
6.1.4 Procedure

In total, 205 participants (see Table 6.5 for condition breakdown) finished the experiment. Figure 6.2 shows the experimental flow of the task. All participants went through an informed consent process. After that, participants reviewed an overview page describing the workflow of the task. Part of the instruction was composed differently based on the group framing conditions. Before reviewing the poster and the narratives, participants filled out a pre-study survey measuring their empathy quotients and attitudes towards negative feedback. Later, participants reviewed the design and read a passage based on narrative empathy conditions. Last, participants composed feedback for the design and filled out a post-study survey regarding their perception of the designer and their attitudes towards negative feedback.

As a member of the company, I had a significant influence on its design. The poster needs to include the logos of the company and the sponsors. The event information and logo also need to be there. Another need was to have photos of the honoree and chairs of the event. They had ties to a local sports team and could help promote the event. My biggest concern with the poster is that it's difficult to determine what type of event it is just by looking at it. The only way to fully understand the event is to go the ticket website link and learn more. I believe that's what my manager hoped would happen. The people at the center would drive traffic to our ticket website. Because the design is Philly themed we went with a picture of the skyline for the background. The company used the original image for previous years' promotional material. So I altered the colors and blurred it to make it somewhat indistinguishable. The edits also allowed the text to stand out and the individuals' headshots to be a focal point. I choose to render all sponsor logos in white. This more consistent color profile also helps to play down that section. The font is Futura, which I like to use because of it is minimal and includes many weights. Weight options are key to creating a hierarchy of importance with text. I'm happy with how it turned out given the circumstances I was working under. The poster was intended to be displayed throughout the city including public commute. It's simple and eye-catching in those settings. I did a good job at updating the original material to the current advertising style.

Table 6.2: Design process narrative: The passage described the purpose of the design, design decisions made in the process, the reason to seek feedback, and how the designer felt about feedback collection.

Figure 6.2: Experimental flow chart of the study. At the beginning of the study, all participants were evenly divided into two groups, one reviewing the ingroup framing instructions and the other control framing. Then all participants filled the same pre-study survey and reviewed the design. Afterward, participants read different passages based on narrative empathy conditions. Participants then provided feedback for the design and filled a post-study survey.



When an object is cool and warm air touches the cool object, the air cools and droplets of water forms on the outside of the object. This is the result of the hot and cold air coming into contact with each other. This water in the air is called water vapor. Water vapor is in the form of a gas. Characteristics of water vapor include it being colorless, odorless, invisible, and has no taste. Humidity is the amount of water vapor in the air. When the in the air turns into a gas it is called evaporation. Water vapor gets into the air day through the process of evaporation. Ocean water, and other bodies of water, is turned into water vapor using the energy from the sun. The molecules of the water is absorbed by the Sun's energy near the surface of the water which then evaporates into the air. The changing of a gas into a liquid is called condensation. An example of condensation is the water which covers a mirror following a hot shower. Another large source of water vapor in the air is when the plants absorb water through their roots and stems into their leaves. The leaves then give off water. The process of plants releasing water into the air is called transpiration. All of the water in the air, whether it is from the world's ocean and other bodies of water, the water on a mirror following a hot shower, or the water a plant releases into the air; it is all called humidity because it is the amount of water vapor in the air.

Table 6.3: Control article

Ingroup Framing Condition
To reward high performance, we offer a bonus opportunity in this HIT. We have grouped the designer of the poster and you as Team Orange and you two will collaborate on this task. After you submit the HIT, the designer of the poster will review your feedback and try to revise the design accordingly. A domain expert will rate the success of your collaboration by evaluating the improvement in the design based on the feedback that you provide. The teams who score in the top 10 will share a cash bonus of 8,4 per team member.
Control Framing Condition
To reward high performance, we offer a bonus opportunity in this HIT. We ask you to complete this task independently. Your feedback should help the designer improve the poster. A domain expert will rate how useful your feedback is by evaluating its potential for helping the designer improve the poster. The participants whose feedback is ranked in the top 10 will earn a cash bonus of $4.

Table 6.4: Participants read different statements about a bonus opportunity based on ingroup framing experimental conditions. Other task instructions also referred to the content creator as "the designer" or "your teammate" based on ingroup framing conditions throughout the task.

6.1.5 Participants

Due to the scale of the experiment, we conducted the experiment on Amazon Mechanical Turk (AMT). To ensure participants were representative of feedback exchange platform users, we adopted a screening process where they answered a question regarding their experience in providing design feedback. To warrant truthful answers, we asked the same screening question again in the post-study survey. Participants with inconsistent answers to these two questions were excluded from the final analysis.

A common issue on AMT was workers' satisfying behaviors [176]. To minimize this behavior, we implemented a series of confirmation checks to ensure participants were performing the tasks as requested throughout the experiment. Participants answered questions about the assigned team label and the content of the narratives. We also added a confirmation check in the pre and post-study surveys about their opinions on an issue unrelated to the experimental manipulation. Participants estimated the popularity of feedback exchange platforms and the participants with notably different answers between the pre and post-study surveys (more than 2 point difference on a 7-point scale) were excluded from the data set. We also excluded participants who repeatedly attempted to skip experimental tasks and participants who spent an unusually long time on the task (two standard deviations higher than the average) from the final analysis.

In the final participant pool, 57% were female, 43% male; 13% were 18-24 years old, 44% 25-34 years old, 25% 35-44 years old, 18% 45-65 years old; 10% had higher school or lower degree, 41% some college or associate degree, 38% bachelor's degree, 12% graduate degree. Regarding the feedback collection experience, 22% received feedback daily, 34% monthly, 32% weekly, and 11% yearly. For the frequency of receiving negative feedback, 11% had never received negative feedback, 19% daily, 28% weekly, 31% monthly, and 12% yearly. We paid each participant $4 ($13.5/hr) upon task completion. Workers who failed attention checks received a payment proportional to HIT duration up to $4. All participants had finished more than 500 HITs on AMT and had a pass rate higher than 98%.

6.1.6 Measurement

The main measures included feedback quality, feedback length, and attitudes towards negative feedback. Prior work has used feedback quality and invested effort (measured by feedback length) to evaluate feedback because they directly impact how much content creators would benefit from the feedback [1, 40, 41, 43, 177, 178, 179, 180]. For feedback quality, we hired two domain experts from UpWork to rate the quality of the collected

feedback separately. We share the instructions feedback providers received with the experts, and asked them to use their own judgment to decide the quality of the feedback. Each expert started with a calibration phase where they rated 30 pieces of randomly sampled feedback. We instructed the experts to use the entire 7-point scale in calibration and rate the rest of the feedback set using the same standard. Both experts gave similar ratings to the feedback (Pearson's r=0.53), and we averaged the ratings as the final measure. We also measured feedback length as a heuristic of the level of invested effort. To examine whether interventions would lead to changes in feedback content and sentiment, one researcher coded the collected feedback at a sentence level using an established feedback schema [181]. Following prior work, we used LIWC to analyze the feedback and examined ratings in relevant categories [182, 183].

We also created an 8-question survey measuring participants' attitudes toward negative feedback to gauge how likely they would take proactive interventions against negative feedback. The survey was crafted based on prior survey work about harassment and bystander intervention [184]. As Table 6.6 shows, the survey had three sections, focusing on participants' attitudes toward the recipients of negative feedback, the occurrence of negative feedback, and their tendency to intervene. At the beginning of the survey, participants reviewed the definition of negative feedback and an example to avoid confusion about later survey questions. To measure changes in their attitudes, participants answered the same set of questions twice before and after the experimental task.

For confirmation checks of the narrative empathy manipulation, participants answered two questions about to what degree they had tried to provide feedback from the designer's perspective (see Table 6.6). Since only participants in the negative experience condition read about the negative feedback the designer had received before, we used perspective taking as a heuristic for their empathy towards the designer. The question was adapted from prior work on interpersonal empathy [185]. For confirmation checks of the ingroup framing manipulation, we used the classic Inclusion of Other in the Self (IOS) scale [186]. We also included an 8-question survey to measure the empathy quotient of the participants and use it as a covariate in the analyses to address individual differences [187].

6.2 RESULTS

Below we report the significant patterns in our results.

	Control Article	Negative Experience	Design Process
Control Framing	34	34	38
Ingroup Framing	34	32	33

Table 6.5: Participant count breakdown by experimental condition. There were 205 participants in total.

General Attitudes Towards Negative feedback Recipients
• I feel very sorry for people when they receive negative feedback.
• I have tender, concerned feelings for people who receive negative feedback.

General Attitudes Towards Negative feedback
• It is evident to me that people who receive negative feedback need support from other members on the same online platform.
• If someone writes negative feedback, people should realize it is a necessary experience for them to grow.
• I think such negative feedback is hurtful and damaging to people.

General Tendency of Intervention
• I feel personally responsible to intervene and offer support to people when they receive negative feedback.
• Even if I am not the one providing the negative feedback, it is still my responsibility to try to discourage others from doing so.
• I believe that my actions can help to reduce the occurrence of negative feedback.

Perspective Taking
• I tried to make my feedback more useful by imagining how the designer of the poster would react to it.
• I was more concerned about whether my feedback would be useful than how the designer of the poster would react to it.

Table 6.6: Questions in the post-study survey. Participants rated their level of agreement for each statement on a scale from 1 to 7 (strongly agree). Except for the two perspective-taking questions, the other eight questions were also asked in the pre-study survey.

6.2.1 All Interventions Increased Feedback Quality

A two-way ANOVA showed narrative empathy improved feedback quality ($F[2, 198]=2.47$, $p=.088$). Participants wrote higher quality feedback in the design process condition ($\mu=3.81$, $sd=1.03$) than in the control article condition ($\mu=3.38$, $sd=1.42$; $p=.072$ after Tukey's HSD adjustment; Cohen's $d=0.35$). Participants in the negative experience condition wrote feedback of similar quality ($\mu=3.56$, $sd=1.07$; adj. $p=.638$; $d=0.14$) in comparison to the control article condition. We also observed a significant interaction between the two factors ($F[2, 198]=4.05$, $p=.019$). Ingroup framing tended to increase the feedback quality in the control article condition ($F[1, 198]=1.85$, $p=.176$; $d=0.65$). There was no significant difference between the ingroup framing ($\mu=3.70$, $sd=1.22$) and the control framing condition ($\mu= 3.48$, $sd=1.16$; adj. $p=.175$; $d=0.19$).

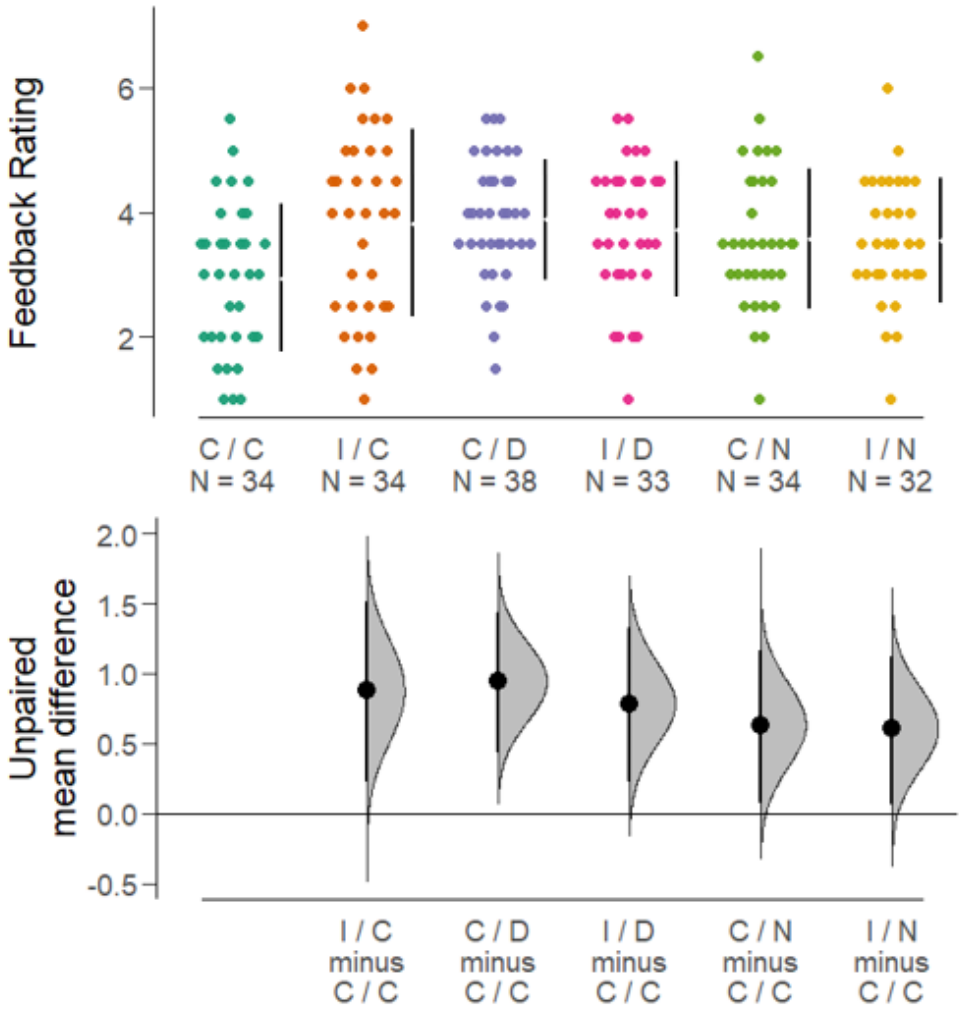
As shown in Figure 6.3, all five intervention conditions resulted in higher feedback ratings compared to the control framing/control article condition. For the five pair-wise planned comparisons, we adjusted the p-value threshold using Holm's Bonferroni method to minimize familywise error [188]. In comparison with the control framing/control article condition ($\mu=2.94$, $sd=1.19$), participants provided significantly higher quality feedback in the control framing/design process ($\mu=3.88$, $sd=0.97$; adj. $p=.002$; $d=0.87$), ingroup framing/design process ($\mu=3.73$, $sd=1.10$; adj. $p=.026$; $d=0.69$), and ingroup framing/control article ($\mu=3.82$, $sd=1.51$; adj. $p=.028$; $d=0.65$) conditions. The ingroup framing/negative experience ($\mu=3.55$, $sd=1.00$; adj. $p=.057$; $d=0.55$) and control framing/negative experience ($\mu=3.57$, $sd=1.14$; adj. $p=.057$; $d=0.54$) conditions also increased feedback quality.

6.2.2 Design Experience and Ingroup Framing Increased Levels of Effort Invested

A two-way ANOVA showed narrative empathy had a main effect ($F[2, 198]=3.04$, $p=.050$). Participants in the design process condition ($\mu=938.0$, $sd=433.4$) wrote longer feedback than the ones in the control article condition ($\mu=800.7$, $sd=433.4$; adj. $p=.17$; $d=0.29$). Negative experience narrative ($\mu=757.1$, $sd=403.6$; adj. $p=.84$; $d=0.09$) had no effect. We also observed a weak interaction effect between the two factors ($F[2, 198]=2.85$, $p=.060$). Ingroup framing led to significant differences when participants read the control article. We report the pairwise differences in detail below. Ingroup framing had no effects ($F[1, 198]=0.53$, $p=.468$; $d=0.10$). There was no significant difference between the ingroup framing ($\mu=858.1$, $sd=445.8$) and the control framing condition ($\mu= 812.0$, $sd=476.6$; adj. $p=.468$).

As Figure 6.4 shows, all five intervention conditions reported longer feedback lengths than the control framing/control article condition. In comparison with the control framing/control

Figure 6.3: Feedback quality across conditions. In comparison with the leftmost control framing/control article condition, both ingroup framing and design process increased ratings of feedback quality. Negative experience had a similar effect. Here we label all group conditions in the format of "A / B". A indicates the group framing condition: Ingroup framing or Control framing, and B indicates the narrative empathy condition: Design process, Negative experience, or Control article.



article condition (μ=668.6, sd=450.1), participants provided significantly longer feedback in the control framing/design process condition (μ=961.0, sd=478.5; adj. p=.047; d=0.63), and notably longer feedback in the ingroup framing/design process (μ=911.5, sd=380.5; adj. p=.080; d=0.58) and ingroup framing/control article condition (μ=932.8, sd=566.6; adj. p=.111; d=0.52) conditions. Participants tended to write more in the control framing/negative experience (μ=788.7, sd=464.9; adj. p=.573; d=0.26) and ingroup framing/negative experience (μ=723.5, sd=330.4; adj. p=.566; d=0.14) conditions.

6.2.3 Feedback Tends to Have More Positive Tones

We performed an LIWC analysis on the collected feedback (see Table 6.7). Following prior work, we used analytical, social, and tone as three main categories because of their relevance to the task [182]. Our results showed feedback across conditions had similar levels of analytical and social ratings. Feedback in all the intervention conditions had higher positive tone than in the control framing/control article condition. We did not observe statistical differences regarding feedback categories and LIWC ratings. One reason might be we conducted the study in a realistic setting without offering strong stimulus to negative feedback as in prior work [5, 182]. Instead, the interventions had significant influences on how participants perceived negative feedback.

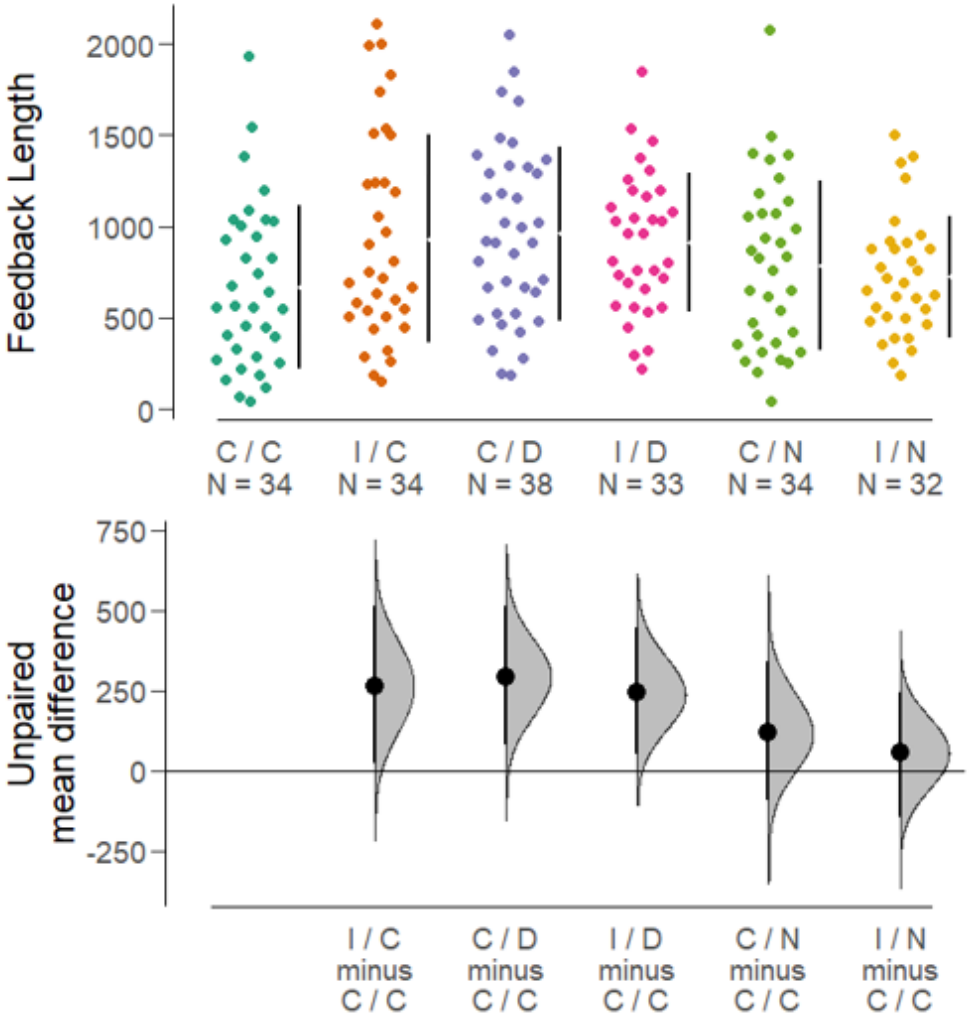
	control article		design process		negative experience	
	cont.	ingr.	cont.	ingr.	cont.	ingr.
analyt.	81.3 (19.1)	79.4 (22.1)	77.3 (18.7)	80.3 (12.6)	77.2 (17.5)	82.5 (16.5)
social	6.3 (3.7)	6.0 (2.6)	6.5 (2.9)	5.7 (2.6)	7.1 (4.4)	6.0 (2.9)
tone	***71.0***	***76.2***	***78.2***	***78.6***	***80.4***	***75.4***
	(26.7)	***(24.4)***	***(21.0)***	***(19.9)***	***(22.8)***	***(24.3)***

Table 6.7: Mean and (standard deviations) for LIWC analysis across experimental conditions. All three output variables range from 0 to 100. Feedback in the intervention conditions had higher positive tone (in bold) than in the control framing/control article condition (in italics). For reference, intervention conditions reported a tone level similar to the one of natural speech (79.29) while the control framing/control article condition's level is slightly lower than the one of Twitter posts (72.24) [189].

6.2.4 No Difference Found Across Conditions Regarding Feedback Categories

After coding the feedback using a scheme developed in prior work [181], we summed the length of feedback written in each feedback category in each experimental condition. Then we

Figure 6.4: Feedback length was used as a heuristic for the effort invested in feedback composition. The leftmost bar represents the control framing/control article condition. Here we observed a similar trend to the one in feedback quality. Ingroup framing and design process increased feedback length; negative experience had a similar pattern but less pronounced.



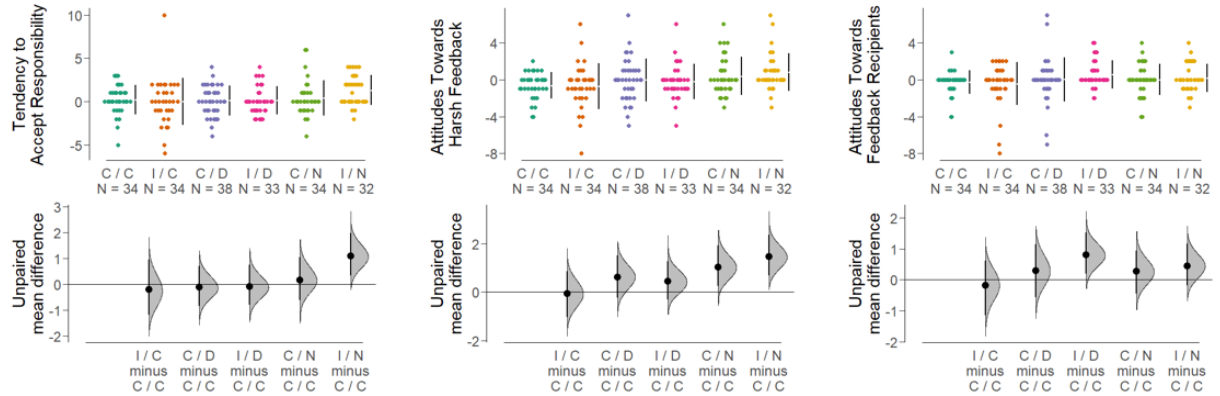
normalized the counts to calculate the ratios. There were no significant differences between the conditions in terms of the ratios of feedback categories. On average, 7.6% of the feedback was positive only, 18.9% positive and specific, 31.6% problem only, 29.4% solution only, 9.4% problem and solution, and 3.1% others. On average, participants expanded each feedback category proportionally when they wrote longer feedback in the intervention conditions.

6.2.5 Negative Experience Led to More Disapproving Stance Against Negative feedback

For participants' pre and post-study differences in attitudes towards negative feedback, we summed the differences into three measures: attitudes towards recipients of negative feedback, attitudes towards negative feedback itself, and tendency to accept responsibility to help.

Narrative empathy had a main effect on participants' tendency to accept responsibility to help ($F[2, 198]=3.12$, $p=.046$; see Figure 6.5). In comparison with the control article condition ($\mu=0.10$, $sd=2.22$), negative experience ($\mu=0.83$, $sd=1.92$; adj. $p=.076$; $d=0.35$) tended to make participants more likely to accept responsibility to help. In particular, the ingroup framing/negative experience condition ($\mu=1.31$, $sd=1.71$) reported significantly stronger tendency in comparison with the control framing/control article condition ($\mu=0.21$, $sd=1.67$; adj. $p=.049$; $d=0.66$).

Figure 6.5: The above charts show how interventions influenced participants' attitudes towards negative feedback and its recipients. Among the three interventions, negative experience was most effective. It led to a significantly more disapproving stance towards negative feedback and made participants significantly more likely to intervene.



Narrative empathy also had a main effect over participants' attitudes towards negative feedback ($F[2, 198]=6.32$, $p=.002$; see Figure 6.5). In comparison with the control article

condition (μ=-0.68, sd=2.00), negative experience (μ=0.59, sd=2.05; adj. p=.001; d=0.63) made participants take a significantly more disapproving stance towards negative feedback. Interestingly, when participants performed the task without any intervention they reported more tolerance of negative feedback (paired t-test p=.013; d=0.15). The experience of providing feedback might make participants more aligned with the perspectives of feedback providers and more lenient with the behaviors.

For attitudes towards recipients of negative feedback, narrative empathy did not have a main effect (F[2, 198]= 2.11, p=.124; see Figure 6.5). But both the design process (μ=0.27, sd=2.02; adj. p=.114; d=0.32) and negative experience (μ=0.09, sd=1.61; adj. p=.343; d=0.26) conditions led to more empathetic feelings towards the recipients than the control article condition (μ=-0.35, sd=1.85). In comparison with the control framing/control article condition (μ=-0.26, sd=1.29), participants in the ingroup framing/design process condition were notably more supportive to the feedback recipients (μ =0.55, sd=1.50; adj. p=.105; d=0.58).

6.2.6 Manipulation Checks of Factors

Ingroup framing had a main effect on group perception (F=5.879; p=.016). Participants in the ingroup framing condition reported significantly higher IOS (μ= 3.45, sd=1.51) than participants in the control framing (μ=2.90, sd=1.76; adj. p=.016; d=0.34). No significant effect was detected regarding participants' perspective-taking. The ingroup framing/negative experience condition (μ=8.34, sd=2.72) tended to encourage more perspective-taking in comparison with the control framing/control article condition (μ=7.03, sd=2.54; adj. p=.235; d=0.50).

6.2.7 Empathy Quotient Caused No Difference

We used participants' empathy quotient as a covariate to control individual differences in personalities. This measure did not influence the significance levels of any analysis. The effects we uncovered in our study were applicable to all participants with various levels of empathy capabilities.

6.3 DISCUSSION

All the proposed interventions were effective at increasing feedback quality and effort invested in the task (measured by feedback length). While all participants received the

same payment and task instructions, ingroup framing increased feedback quality by 30% and feedback length by 40%; the design process narrative by 32% and 44%; and the negative experience condition by 21% and 18%, respectively. Regarding attitudes towards negative feedback and its recipients, participants in the negative experience condition reported a more negative attitude toward negative feedback and were more likely to accept responsibility to intervene when negative feedback is observed relative to participants in the other conditions. Participants in the design narrative condition became more supportive to the recipients of negative feedback.

Both the ingroup framing and design process narrative conditions achieved similar effects, i.e., increasing feedback quality and effort invested in feedback composition. Platform designers should prefer to use ingroup framing when this intervention is applicable, as it imposes minimal, if any, additional task load on either the recipient or provider of feedback. In our study, we tested two common ways of ingroup framing: group interdependency and group labeling. For interdependency, platform designers could promise various rewards for collaboration. While monetary rewards may not always be suitable, platform designers could use reward points or badges as alternatives [190, 191]. Future work could also test platforms where successful collaboration helps users to earn privileges related to feedback exchange, such as longer exposure in the content feed so they could collect more feedback, or ability to invite experienced members to provide expert reviews.

For the implementation of team labeling, platform designers could ask users to generate their own names or select among provided options after forming teams on-demand [192]. Alternatively, platform designers could consider using labels based on an existing relationship, such as shared interests in a design genre/style, similar years of experience, or adjacent time of joining the platform. Future work could also test encouraging feedback providers to proactively form groups with designers. In our study, we tested short-term ingroup framing for one design-feedback cycle. Future work could examine long-term framing spanning multiple projects and how the effects of the framing might change over time. Future work could also explore what proportions of the observed effects could be attributed to interdependency and team labeling respectively.

Ingroup framing may be inapplicable in some scenarios. For example, when it is critical for feedback providers to provide an objective analysis of the work, an ingroup framing may bias their evaluation of the work. In these scenarios, the design process intervention may serve as an alternative. Platform designers could provide guidelines and templates to help content creators to write an effective design process narrative. Since some content creators may be unwilling to invest the effort, platform designers could offer this as a suggestion and highlight the benefits of receiving higher quality feedback. Prior work has explored

scaffolding processes that help users to craft effective help-seeking emails [193]. Future work could explore similar scaffolding that makes it easier for content creators to write an effective narrative. Researchers have also explored recording design processes via design editor add-ons and re-creating the process using action logs [194, 195]. Afterward, a content creator could annotate key frames to quickly compose a design process narrative.

Negative experience narrative is most effective at encouraging feedback providers to take a more disapproving stance against negative feedback on the platform. This is particularly important as our results showed participants were more tolerant of negative feedback after performing the task in the control framing/control article condition. The experience of providing feedback might have made them more inclined to justify negative feedback. Negative experience helps to reverse this trend. Platform designers could selectively present this feature if negative feedback reaches an undesirable level. Since asking the designer to write about past negative feedback incurs additional work, platform designers could consider alternative methods to mitigate the costs. One way is to use negative feedback the content creator received previously to showcase the negative experience. Meanwhile, content creators could choose to paraphrase the exchange and add their emotional responses to make the intervention more effective. Previous work shows negative valence posts may make users feel this type of content is acceptable and further incite more posts of similar valence [5]. Content creator's comments may negate this influence by conveying how such feedback is undesirable. Platform designers could also consider creating a pool of negative experience narratives and present them during an onboarding process of the community. We also observed that the negative experience narrative condition stimulated participants to accept responsibility to intervene when negative feedback occurred. Future work should explore if this reported attitude translates into intervening action beneficial for the feedback exchange community.

We observed mixed results regarding the interaction of the interventions. The ingroup framing/negative experience condition led to the highest level of perspective-taking, a core aspect of empathy. Platform designers could use these two conditions together if empathy towards the designer is most important. Such a scenario may occur if the affective states of a content creator had recently been affected by unfavorable interactions, such as receiving negative feedback. On the other hand, using ingroup framing together with other narrative empathy conditions did not further enhance the effects of the narratives in terms of feedback quality and invested effort level. Platform designers could use a single intervention to maximize these measures while minimizing the overhead in implementation. One possible explanation of this pattern is diminishing marginal utility. With the same payment, a second intervention might not be enough to elicit meaningfully more effort in the task. Future

work could test these interventions with participants under different incentive schemes and examine whether the combinations of the interventions could indeed lead to even higher feedback quality and effort levels.

We measured participants' empathy quotient in this study to control individual differences in their capacity to feel empathy towards the content creator. The results show empathy quotient had no main effects over the study measures. However, future work should keep measuring individual differences when evaluating other interventions. There may also be individual differences in terms of their susceptibility to various empathy arousal interventions. While we examined the proposed interventions in a between subject study, future work could conduct within subject studies to examine participants' individual differences in intervention susceptibility.

6.4 LIMITATIONS

We used a single poster design in our experiment. Future work should evaluate the generalizability of our results to other genres of creative work. We may observe different patterns for artistic expressions with more abstract goals, where the standard for high-quality feedback is more subjective and less clear. Also, for designs that require a substantially longer time to review, such as a feature movie or a book, our interventions may have different effect sizes. Most participants in our experiment had a moderate amount of experience in design. Domain experts with extensive experience may also react differently to the interventions.

CHAPTER 7: GENERAL DISCUSSION

In this section, I would like to compare all proposed methods in this thesis and evaluate their effectiveness over the collected measures.

Throughout this thesis, I have examined three categories of interventions, self-directed coping activities (Chapter 4), valence-based ordering of feedback (Chapter 5), and empathy arousal (Chapter 6). All three experiments collected both behavioral and attitudinal measurements from participants. For behavioral measures, regardless whether targeting at the feedback provider or the receiver end, I measured both the level of effort invested (estimated by feedback length or edit distance in revision) and the quality of the work (feedback or essay quality rated by domain experts). Among the interventions, empathy arousal is the most effective one at encouraging participants to invest more effort and thus lead to higher quality work. Both ingroup framing and narrative empathy interventions (e.g. design process) resulted in significant improvements. The other two categories of interventions did not have any main effects, although valence-based ordering significantly increased participants' affective states, which were correlated with revision extents.

For attitudinal measures, I measured participants' perception of the feedback and its providers on the content creator end, and their attitudes towards negative feedback and willingness to help the content creator on the feedback provider end. All three categories of interventions were effective at casting positive influences over these attitudinal measures. Among the interventions, negative experience narrative and valence-based reordering were most effective. The former led to a significantly more disapproving stance against negative feedback and more likely to intervene in the face of negative feedback; the latter significantly improved participants' self-reported ratings of enthusiasm, excitement, and happiness. One self-directed coping activity — distraction — had similar effects over affective states to a lesser degree. Other interventions had no effects.

Overall, there is no silver bullet addressing the problem of negative feedback, i.e., no single intervention outperforms all the others over all measures. Because of these, platform designers should consider adopting at least two interventions at the same time, one on the feedback provider end and one on the receiver end. Platform designers should also determine their own priority of different values and select the most appropriate intervention on both ends based on their preference. Multiple interventions could be combined and adopted at the same time if the platform would like to improve multiple measures. But at the same time, I would like to caution against using too many interventions as this risks incurring too much additional work to users and thus impairs the overall efficiency of the feedback

exchange process.

Besides the efficacy of the interventions, another important factor to consider here is how much additional work the participants had to perform for the intervention. From this perspective, both the ingroup framing and valence-based ordering require minimal amount of work, as they could be incorporated into the existing feedback exchange process organically without the need for adding additional steps. Surprisingly, these undemanding manipulations notably outperformed interventions that require more effort, such as self-directed coping activities, for many measures discussed above. At a deeper level, it appears urging users to view the feedback exchange process from a new perspective, while giving users more freedom in feedback composition, may be more effective than cajoling them into following specific guidelines. Platform designers should have long term visions regarding what kind of culture they should cultivate within the community and what the ideal relationships between feedback providers and content creators should be.

In my experiments, I collected a set of measures that helped to evaluate the efficacy of the interventions. Now looking back, some other measures may also be useful for comparing the interventions and evaluate the feasibility of them on real-world platforms. One is how effective these interventions are in the long term if the users perform them routinely. One possibility may be users have internalized the interventions and the relevant measures would move in favorable directions without using interventions. But at the same time, another possibility is users getting used to seeing the interventions and start gaming the system, which made the interventions less effective over time. Future work could help to answer this question. Another useful measure to collect is the cognitive load required by the interventions. Here I have timestamped how long it took to finish specific interventions, but there is no quantitative measure regarding how mentally taxing the intervention is. This measure can help platform designers to choose interventions as users generally prefer activities with lighter cognitive loads. Future experiments could also ask users about how much they like to see the intervention being implemented on platforms, after informing users about the resulted benefits. User preferences can inform platform designers' decision when selecting coping interventions.

In my thesis, we evaluated all interventions in experimental settings. It is still unknown how exactly users will react to these mechanisms on real-world platforms. Future work may consider examining these interventions in field studies by implementing these interventions as add-ons to existing platforms. I foresee a few factors may affect the effectiveness of the interventions and the influence can happen in both favorable and unfavorable directions. These factors include different stages of a design project (early stage users may be more open to feedback), different categories of creative projects (more costly ones may be more likely

collect more feedback and iterate more), and budget or time constraints (fewer iterations because of these limits). Future empirical work could help to shed more light upon these issues.

CHAPTER 8: FUTURE WORK

For both valence-based order and the coping activities, we tested the interventions on three pieces of feedback because this was sufficient to answer the research questions being asked. Future work should test the generalizability of our results to larger feedback sets. When evaluating the valence-based order mechanism, we show presenting positive feedback first mitigates the influences of negative feedback. For other experiments, we used neutral valence feedback to avoid confounding effects. Researchers could also examine the effectiveness of the interventions when each piece of feedback has a mix of positive and negative phrasings. Participants in our experiments were novices in design. Future work is needed to test how our results would generalize to users with different levels of design and domain expertise, as more experienced users may already have developed their own mechanisms to cope with negative feedback. Another issue that needs further testing is how our findings transfer to other types of creative work that requires significantly more effort, such as video production, or ones that require more logical reasoning, such as programming. We hope our study can serve as a starting point for future work that continues to test other activities that may improve people's resilience to negative feedback, such as relaxation [196], music therapy [197], and social support [198], and activities that induce empathy via perspective-taking, such as role-play games [199], educational videos [200], and drama [201].

Another direction to explore is long term interventions. In our work, we showed how empathy arousal methods, including narrative empathy and ingroup framing, promote positive feedback and increase the feedback quality. Future work could explore how these interventions shape the behaviors within the community in the long term. Researchers could explore other community building practices, such as providing a discourse space [202], forming user organizations [203], matching users with similar interests [203], and examine how improving the relationship between content creators and feedback providers influences feedback exchange behaviors. Future work could also examine whether frequently performing the proposed interventions could enable content creators to naturally incorporate them as a integral step of their design process. In this case, content creators may become more resilient to negative feedback over time while lowering their dependency on the interventions. On the other end, future work could also explore ways to educate feedback providers to provide more positive and constructive feedback using long term interventions. Platform designers may consider offer rubrics and guidelines in a procedural way. New community members may start by following guidelines that are easy for them to follow. Later, they may receive additional training as they gain experience on the platform. An alternative way is to offer a

scaffolding at the beginning, and reduce support while giving more freedom to users as they gain experience.

This dissertation explores interventions to reduce the effects of negative feedback that a content creator might receive on their creative projects. However, it is equally important to help users build resilience to negative feedback, especially since the interventions explored in this dissertation might not be available in all instances where a creator receives feedback. These two perspectives are not necessarily in tension, for example, the interventions explored in this thesis could be used to help a novice increase their resilience to negative feedback. One common practice in design education is to develop students' resilience towards harsh criticism throughout their degree program [204]. Researchers could explore interventions that gradually build up users' resilience towards negative feedback over time. As such processes often require content creators to routinely experience negative feedback, it is crucial to evaluate the cost of resilience-building interventions in terms of their short and long-term effects on the creators' affective states and the percentage of creators that become disheartened in the process. How to selectively apply the interventions to build resilience is an open question for future work.

In Chapter 6, I have explored how monetary incentives may create interdependency between content creators and feedback providers and develop an in-group framing. Future work could experiment with introducing such incentive structures to real-world platforms and examine the change in the frequency and quality of the exchanged feedback. Researchers could also explore the feasibility of creating ingroup framing using gamified community points. Such mechanisms may be more viable on platforms that have already embraced gamification. Future work could also explore using a mixture of incentives include both fiat currency and community points. Feedback providers may receive different types of rewards based on their contribution type. For example, platforms could reward them fiat currency for providing feedback and community points for effective teamwork. Administrators of paid feedback exchanged platforms could explore these options without worrying about significant disturbance to the current microeconomics of the platform.

Researchers could continue to conduct additional experiments situated within the design space described in Table 1.1. Future work could focus on the table cells where have been less explored by researchers. For example, currently feedback receivers have been relying on white and blacklists to control who could provide feedback. Another direction is to explore softer boundaries, such as blocking users after detected aggressive actions did not stop after a period of time. Community administrators have relied on ad hoc censoring to remove negative feedback and block that user from further participation. Future work could explore more responsive mechanisms, such as banning users only when malicious actions continue.

On existing platforms, feedback receivers usually have little control over who is eligible for seeking feedback from the community. Future work may explore forming community committees that enforce a quality standard for feedback seeking posts and users who fall short will no longer be allowed to seek additional feedback.

Future work may also explore interventions that combine multiple steps and roles in Table 1.1. For example, researchers could explore smart routing algorithms that determine both who provides and who receives feedback. The algorithm could match users based on their preference for design styles and the needs for feedback at different design stages. Future work could also give feedback providers more control over how the feedback will be presented and consumed in addition to their control over feedback composition. Feedback providers may reserve more critical feedback for experienced users and conceal the feedback from novice users. Future work could also explore mechanisms that let feedback provider and receivers to take a more active role in the mechanism design, giving them more control over critical settings regarding feedback exchange, such as design and feedback ordering, content censoring, white/blacklists, etc.

In my work, I examined the proposed interventions within the space of design critique. Future work could explore the effectiveness of these interventions in other domains, such as schoolwork critique in design education, performance review in corporate environments, etc.

CHAPTER 9: CONCLUSION

This thesis developed and tested techniques for each stage of the feedback exchange process – including the generation, presentation, and consumption of feedback – to mitigate the occurrence of negative feedback and its effects. Specifically, the contributions are: 1) feedback providers write better feedback if they read an empathy-arousing narratives prior to the task; 2) recipients are less affected by negative feedback if it is read after feedback that has more positive sentiment in that same collection; and 3) recipients can further mitigate the effects of negative feedback by performing self-directed coping activities such as self-affirmation, distraction, and expressive writing. After examining and quantifying the effects of negative feedback and exploring various methods to address the problem, we hope the new approaches proposed in my thesis can encourage future researchers in the HCI community to continue exploring other novel mechanisms. I hope the contributions of this thesis progress toward a future where content creators can receive useful and supportive feedback online and, in cases which fall short of this ideal, can perform or experience techniques to minimize the effects of the negative feedback.

REFERENCES

[1] Y.-C. G. Yen, S. P. Dow, E. Gerber, and B. P. Bailey, "Social Network, Web Forum, or Task Market?" in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16*. New York, New York, USA: ACM Press, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2901790.2901820 pp. 773–784.

[2] J. Suler, "The Online Disinhibition Effect," *CyberPsychology & Behavior*, vol. 7, no. 3, pp. 321–326, jun 2004. [Online]. Available: http://www.liebertonline.com/doi/abs/10.1089/1094931041291295

[3] M. Samory and E. Peserico, "Sizing Up the Troll," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. New York, New York, USA: ACM Press, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3025453.3026007 pp. 6943–6947.

[4] R. N. Landers and J. W. Lounsbury, "An investigation of Big Five and narrow personality traits in relation to Internet usage," *Computers in Human Behavior*, vol. 22, no. 2, pp. 283–293, mar 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0747563204001128

[5] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone Can Become a Troll," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. New York, New York, USA: ACM Press, 2017. [Online]. Available: http://www.cs.cornell.edu/{~}cristian/Anyone{_}Can{_}Become{_}a{_}Troll{_}files/anyone{_}can{_}become{_}a{_}troll.pdfhttp://dl.acm.org/citation.cfm?doid=2998181.2998213 pp. 1217–1230.

[6] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, "Bad is stronger than good." *Review of General Psychology*, vol. 5, no. 4, pp. 323–370, 2001. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/1089-2680.5.4.323

[7] A. Aula and V. Surakka, "Auditory Emotional Feedback Facilitates Human-Computer Interaction," in *People and Computers XVI - Memorable Yet Invisible*. London: Springer London, 2002, vol. XVI, pp. 337–349. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124058651000066http://linkinghub.elsevier.com/retrieve/pii/B9780124058651000066http://link.springer.com/10.1007/978-1-4471-0105-5{_}20

[8] P. Cairns, P. Pandab, and C. Power, "The influence of emotion on number entry errors," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, vol. 137, no. 4. New York, New York, USA: ACM Press, 2014. [Online]. Available: http://scholar.harvard.edu/dtingley/files/emotionmanipulationm11.pdfhttp://dl.acm.org/citation.cfm?doid=2556288.2557065 pp. 2293–2296.

[9] P. Resnick and R. Zeckhauser, "Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system," *Advances in Applied Microeconomics*, vol. 11, pp. 127–157, 2002.

[10] C. Lampe and P. Resnick, "Slash(dot) and burn," in *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, vol. 6, no. 1. New York, New York, USA: ACM Press, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?doid=985692.985761 pp. 543–550.

[11] Ai-Mei Chang, P. Kannan, and A. Whinston, "Electronic communities as intermediaries: the issues and economics," in *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers.* IEEE Comput. Soc. [Online]. Available: http://ieeexplore.ieee.org/document/772942/ p. 10.

[12] A. Arnt and S. Zilberstein, "Learning to perform moderation in online forums," in *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003).* IEEE, 2003, pp. 637–641.

[13] A. Yuan, K. Luther, M. Krause, S. I. Vennix, S. P. Dow, and B. Hartmann, "Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16.* New York, New York, USA: ACM Press, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2818048.2819953 pp. 1003–1015.

[14] M. D. Greenberg, M. W. Easterday, and E. M. Gerber, "Critiki : A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers," in *C&C*, 2015.

[15] A. Xu, S.-w. Huang, and B. Bailey, "Voyant," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14.* New York, New York, USA: ACM Press, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2531602.2531604 pp. 1433–1444.

[16] C. M. Mascaro, A. Novak, and S. Goggins, "Shepherding and censorship: Discourse management in the tea party patriots facebook group," in *2012 45th Hawaii International Conference on System Sciences.* IEEE, 2012, pp. 2563–2572.

[17] K. Luther, J.-l. Tolentino, W. Wu, A. Pavel, B. P. Bailey, M. Agrawala, B. Hartmann, and S. P. Dow, "Structuring, Aggregating, and Evaluating Crowdsourced Design Critique," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. New York, New York, USA: ACM Press, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2675133.2675283 pp. 473–485.

[18] C. D. Hunter, "Social Impacts," *Social Science Computer Review*, vol. 18, no. 2, pp. 214–222, may 2000. [Online]. Available: http://journals.sagepub.com/doi/10.1177/089443930001800209

[19] J. Hedgcock and N. Lefkowitz, "Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing," *Journal of Second Language Writing*, vol. 3, no. 2, pp. 141–163, may 1994. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/1060374394900124

[20] "How Reddit Ranking Algorithms Work," pp. https://medium.com/hacking–and–gonzo/how–reddit–ra.

[21] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The Bag of Communities," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. New York, New York, USA: ACM Press, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3025453.3026018 pp. 3175–3187.

[22] "An Introduction to Ranking Algorithms Seen on Social News Aggregators," pp. https://coderwall.com/p/cacyhw/an–introduction–to–ranking–algorithms–seen–on–social–news–aggregators.

[23] J. Rzeszotarski and A. Kittur, "CrowdScape," in *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*. New York, New York, USA: ACM Press, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2380116.2380125 p. 55.

[24] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd," in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. New York, New York, USA: ACM Press, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2047199http://dl.acm.org/citation.cfm?doid=2047196.2047199 p. 13.

[25] S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab, "Searching for safety online: Managing" trolling" in a feminist forum," *The information society*, vol. 18, no. 5, pp. 371–384, 2002.

[26] K. Yatani, M. Novati, A. Trusty, and K. N. Truong, "Review spotlight," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. New York, New York, USA: ACM Press, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1978942.1979167http://dl.acm.org/citation.cfm?doid=1978942.1979167 p. 1541.

[27] T. T. D. T. Nguyen, T. Garncarz, F. Ng, L. A. Dabbish, and S. P. Dow, "Fruitful Feedback," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. New York, New York, USA: ACM Press, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2998181.2998319 pp. 1024–1034.

[28] "Commenting - Reddit.Com," p. https://www.reddit.com/wiki/commenting#wiki_postin.

[29] G. L. Cohen and D. K. Sherman, "The Psychology of Change: Self-Affirmation and Social Psychological Intervention," *Annual Review of Psychology*, vol. 65, no. 1, pp. 333–371, jan 2014. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-psych-010213-115137http://www.ncbi.nlm.nih.gov/pubmed/24405362

[30] C. R. Critcher, D. Dunning, and D. a. Armor, "When Self-Affirmations Reduce Defensiveness: Timing Is Key," *Personality and Social Psychology Bulletin*, vol. 36, no. 7, pp. 947–959, jul 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0146167210369557

[31] D. K. Sherman and G. L. Cohen, "The Psychology of Self-defense: Self-Affirmation Theory," in *Advances in Experimental Social Psychology*, 2006, vol. 38, no. 06, pp. 183–242. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0065260106380045

[32] S. J. Lepore and J. M. Smyth, Eds., *The writing cure: How expressive writing promotes health and emotional well-being.* Washington: American Psychological Association, 2002. [Online]. Available: http://content.apa.org/books/10451-000

[33] G. Ramirez and S. L. Beilock, "Writing About Testing Worries Boosts Exam Performance in the Classroom," *Science*, vol. 331, no. 6014, pp. 211–213, jan 2011. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1199427

[34] C. T. Chambers, A. Taddio, L. S. Uman, and C. M. McMurtry, "Psychological interventions for reducing pain and distress during routine childhood immunizations: A systematic review," *Clinical Therapeutics*, vol. 31, no. SUPPL. 2, pp. S77–S103, jan 2009. [Online]. Available: http://dx.doi.org/10.1016/j.clinthera.2009.07.023http://linkinghub.elsevier.com/retrieve/pii/S0149291809002628

[35] B. F. Hudson, J. Ogden, and M. S. Whiteley, "Randomized controlled trial to compare the effect of simple distraction interventions on pain and anxiety experienced during conscious surgery," *European Journal of Pain (United Kingdom)*, vol. 19, no. 10, pp. 1447–1455, nov 2015. [Online]. Available: http://doi.wiley.com/10.1002/ejp.675

[36] B. Baird, J. Smallwood, M. D. Mrazek, J. W. Y. Kam, M. S. Franklin, and J. W. Schooler, "Inspired by Distraction," *Psychological Science*, vol. 23, no. 10, pp. 1117–1122, oct 2012. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0956797612446024

[37] C. D. Batson and E. al, "Influence of self-reported distress and empathy on egoistic versus altruistic motivation to help." *Journal of Personality and Social Psychology*, vol. 45, no. 3, pp. 706–718, 1983. [Online]. Available: http://content.apa.org/journals/psp/45/3/706

[38] L. M. Brown, M. M. Bradley, and P. J. Lang, "Affective reactions to pictures of ingroup and outgroup members," *Biological Psychology*, vol. 71, no. 3, pp. 303–311, mar 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0301051105000918

[39] I. Rae, L. Takayama, and B. Mutlu, "One of the gang," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. New York, New York, USA: ACM Press, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2207676.2208723{\%}5Cnhttp://dl.acm.org/citation.cfm?id=2208723http://dl.acm.org/citation.cfm?doid=2207676.2208723 p. 3091.

[40] Y. W. Wu and B. P. Bailey, "Soften the pain, increase the gain: Enhancing users' resilience to negative valence feedback," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–20, 2018.

[41] Y. W. Wu and B. P. Bailey, "Bitter Sweet or Sweet Bitter?" in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition - C&C '17*. New York, New York, USA: ACM Press, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3059454.3059458 pp. 137–147.

[42] Y. W. Wu and B. P. Bailey, "Better feedback from nicer people: Narrative empathy and ingroup framing improve feedback exchange," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW, 2020.

[43] Y. W. Wu and B. P. Bailey, "Novices Who Focused or Experts Who Didn't?" in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. New York, New York, USA: ACM Press, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2858036.2858330 pp. 4086–4097.

[44] P. Baker, "Moral panic and alternative identity construction in Usenet," *Journal of Computer-Mediated Communication*, vol. 7, no. 1, pp. 0–0, jun 2006. [Online]. Available: http://doi.wiley.com/10.1111/j.1083-6101.2001.tb00136.x

[45] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, "Trolls just want to have fun," *Personality and Individual Differences*, vol. 67, pp. 97–102, sep 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0191886914000324

[46] Y. Cheng, C. P. Lin, H. L. Liu, Y. Y. Hsu, K. E. Lim, D. Hung, and J. Decety, "Expertise Modulates the Perception of Pain in Others," *Current Biology*, vol. 17, no. 19, pp. 1708–1713, 2007.

[47] D. Ferris and B. Roberts, "Error feedback in L2 writing classes," *Journal of Second Language Writing*, vol. 10, no. 3, pp. 161–184, aug 2001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S106037430100039X

[48] E. H. Mory, "Feedback Research Revisited," in *Handbook of research on educational communications and technology*, 2nd ed., D. H. Jonassen, Ed. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2004, ch. 29, pp. 745–783.

[49] D. R. Sadler, "Formative assessment and the design of instructional systems," *Instructional Science*, vol. 18, no. 2, pp. 119–144, jun 1989. [Online]. Available: http://link.springer.com/10.1007/BF00117714

[50] O. T. J. ten Cate, "Why receiving feedback collides with self determination," *Advances in Health Sciences Education*, vol. 18, no. 4, pp. 845–849, 2013.

[51] A. Xu and B. Bailey, "What do you think?" in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. New York, New York, USA: ACM Press, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2145204.2145252 p. 295.

[52] H. Zhu, A. Zhang, J. He, R. E. Kraut, and A. Kittur, "Effects of peer feedback on contribution," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. New York, New York, USA: ACM Press, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2470654.2481311 p. 2253.

[53] P. Kollock and M. Smith, "Managing the virtual commons," 1996, p. 109. [Online]. Available: https://benjamins.com/catalog/pbns.39.10kol

[54] J. Preece, *Online Communities: Designing Usability and Supporting Sociability*. John Wiley & Sons, Inc., dec 2000.

[55] C. Dellarocas, "Designing reputation systems for the social web," *Boston U. School of Management Research Paper*, no. 2010-18, 2010.

[56] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, dec 2010. [Online]. Available: http://link.springer.com/10.1007/s13042-010-0001-0

[57] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," no. 2, pp. 1–5, feb 2014. [Online]. Available: http://arxiv.org/abs/1402.3722

[58] X. Zhang and Y. LeCun, "Text Understanding from Scratch," pp. 1–9, feb 2015. [Online]. Available: http://arxiv.org/abs/1502.01710

[59] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," no. NeurIPS, pp. 1–18, jun 2019. [Online]. Available: http://arxiv.org/abs/1906.08237

[60] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," mar 2019. [Online]. Available: http://arxiv.org/abs/1903.09588

[61] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 328–339, jan 2018. [Online]. Available: http://arxiv.org/abs/1801.06146

[62] R. Johnson and T. Zhang, "Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings," *33rd International Conference on Machine Learning, ICML 2016*, vol. 2, pp. 794–802, feb 2016. [Online]. Available: http://arxiv.org/abs/1602.02373

[63] R. Johnson and T. Zhang, "Deep Pyramid Convolutional Neural Networks for Text Categorization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. [Online]. Available: http://aclweb.org/anthology/P17-1052 pp. 562–570.

[64] X. Liu, P. He, W. Chen, and J. Gao, "Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding," 2019. [Online]. Available: http://arxiv.org/abs/1904.09482

[65] X. Liu, P. He, W. Chen, and J. Gao, "Multi-Task Deep Neural Networks for Natural Language Understanding," pp. 4487–4496, 2019.

[66] A. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré, "Training Complex Models with Multi-Task Weak Supervision," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4763–4771, 2019.

[67] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, p. 82, apr 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2436256.2436274

[68] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, "SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods," oct 2016. [Online]. Available: http://arxiv.org/abs/1610.03771

[69] F. Liu, T. Cohn, and T. Baldwin, "Recurrent Entity Networks with Delayed Memory Update for Targeted Aspect-Based Sentiment Analysis," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018. [Online]. Available: http://aclweb.org/anthology/N18-2045 pp. 278–283.

[70] C. M. Steele, "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," in *Advances in Experimental Social Psychology*, 1988, vol. 21, pp. 261–302. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0065260108602294http://linkinghub.elsevier.com/retrieve/pii/S0065260108602294

[71] D. K. Sherman and G. L. Cohen, "Accepting Threatening Information: Self–Affirmation and the Reduction of Defensive Biases," *Current Directions in Psychological Science*, vol. 11, no. 4, pp. 119–123, aug 2002. [Online]. Available: http://journals.sagepub.com/doi/10.1111/1467-8721.00182

[72] M. B. Reed and L. G. Aspinwall, "Self-affirmation reduces biased processing of health-risk information," *Motivation and Emotion*, vol. 22, no. 2, pp. 99–132, 1998.

[73] P. R. Harris and T. Epton, "The Impact of Self-Affirmation on Health Cognition, Health Behaviour and Other Health-Related Responses: A Narrative Review," *Social and Personality Psychology Compass*, vol. 3, no. 6, pp. 962–978, 2009. [Online]. Available: http://dx.doi.org/10.1111/j.1751-9004.2009.00233.x{\%}5Cnhttp://doi.wiley.com/10.1111/j.1751-9004.2009.00233.x

[74] C. L. Toma, "Affirming the self through online profiles," in *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, vol. 62, no. 2. New York, New York, USA: ACM Press, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1753326.1753588 p. 1749.

[75] J. W. Pennebaker, "Writing About Emotional Experiences as a Therapeutic Process," *Psychological Science*, vol. 8, no. 3, pp. 162–166, 1997.

[76] R. Lazarus, "From Psychological Stress to the Emotions: A History of Changing Outlooks," *Annual Review of Psychology*, vol. 44, no. 1, pp. 1–21, jan 1993. [Online]. Available: http://psych.annualreviews.org/cgi/doi/10.1146/annurev.psych.44.1.1

[77] W. J. McKEACHIE, D. POLLIE, and J. SPEISMAN, "Relieving anxiety in classroom examinations." *Journal of abnormal psychology*, vol. 50, no. 1, pp. 93–8, jan 1955. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/13232935

[78] H. Duijnhouwer, F. J. Prins, and K. M. Stokking, "Feedback providing improvement strategies and reflection on feedback use: Effects on students' writing motivation, process, and performance," *Learning and Instruction*, vol. 22, no. 3, pp. 171–184, jun 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0959475211000831

[79] P. C. Broderick, "Mindfulness and Coping with Dysphoric Mood: Contrasts with Rumination and Distraction," *Cognitive Therapy and Research*, vol. 29, no. 5, pp. 501–510, oct 2005. [Online]. Available: http://link.springer.com/10.1007/s10608-005-3888-0

[80] B. A. Thyer, J. D. Papsdorf, D. P. Himle, B. S. McCann, S. Caldwell, and M. Wickert, "In vivo distraction-coping in the treatment of test anxiety." *Journal of clinical psychology*, vol. 37, no. 4, pp. 754–64, oct 1981. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7309864

[81] A. de Rooij and S. Jones, "Mood and creativity," in *Proceedings of the 9th ACM Conference on Creativity & Cognition - C&C '13*. New York, New York, USA: ACM Press, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2466627.2466658 p. 362.

[82] M. Baas, C. K. W. De Dreu, and B. A. Nijstad, "A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus?" *Psychological Bulletin*, vol. 134, no. 6, pp. 779–806, 2008.

[83] J. Zhou, "Feedback valence, feedback style, task autonomy, and achievement orientation: Interactive effects on creative performance." *Journal of Applied Psychology*, vol. 83, no. 2, pp. 261–276, 1998. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.83.2.261

[84] A. de Rooij, P. J. Corr, and S. Jones, "Emotion and Creativity: Hacking into Cognitive Appraisal Processes to Augment Creative Ideation," in *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition - C&C '15*. New York, New York, USA: ACM Press, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2757226.2757227 pp. 265–274.

[85] J. Parkes, S. Abercrombie, and T. McCarty, "Feedback sandwiches affect perceptions but not performance," *Advances in Health Sciences Education*, vol. 18, no. 3, pp. 397–407, 2013.

[86] J. L. Bienstock, N. T. Katz, S. M. Cox, N. Hueppchen, S. Erickson, and E. E. Puscheck, "To the point: medical education reviews-providing feedback," *American Journal of Obstetrics and Gynecology*, vol. 196, no. 6, pp. 508–513, 2007.

[87] E. a. Hesketh and J. M. Laidlaw, "Developing the teaching instinct," *Medical teacher*, vol. 24, no. 3, pp. 245–348, 2002.

[88] M. C. Porte, G. Xeroulis, R. K. Reznick, and A. Dubrowski, "Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills," *The American Journal of Surgery*, vol. 193, no. 1, pp. 105–110, jan 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0002961006006398

[89] D. Fedor, "Performance improvement efforts in response to negative feedback: the roles of source power and recipient self-esteem," *Journal of Management*, vol. 27, no. 1, pp. 79–97, feb 2001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0149206300000878

[90] R. H. Haswell, "NCTE/CCCC's recent war on scholarship," *Written Communication*, vol. 22, no. 2, pp. 198–223, 2005.

[91] C. C. Bracken, L. W. Jeffres, and K. A. Neuendorf, "Criticism or Praise? The Impact of Verbal versus Text-Only Computer Feedback on Social Presence, Intrinsic Motivation, and Recall," *CyberPsychology & Behavior*, vol. 7, no. 3, pp. 349–357, 2004. [Online]. Available: http://www.liebertonline.com/doi/abs/10.1089/1094931041291358

[92] N. Eisenberg and R. A. Fabes, "Empathy: Conceptualization, measurement, and relation to prosocial behavior," *Motivation and Emotion*, vol. 14, no. 2, pp. 131–149, jun 1990. [Online]. Available: https://link.springer.com/content/pdf/10.1007{\%}2FBF00991640.pdfhttp://link.springer.com/10.1007/BF00991640

[93] F. Herrera, J. Bailenson, E. Weisz, E. Ogle, and J. Zaki, "Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking," *PLOS ONE*, vol. 13, no. 10, p. e0204494, oct 2018. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0204494

[94] T. K. Vescio, G. B. Sechrist, and M. P. Paolucci, "Perspective taking and prejudice reduction: the mediational role of empathy arousal and situational attributions," *European Journal of Social Psychology*, vol. 33, no. 4, pp. 455–472, jul 2003. [Online]. Available: http://doi.wiley.com/10.1002/ejsp.163

[95] S. Gair, "Inducing Empathy: Pondering Students' (In)Ability to Empathize With an Aboriginal Man's Lament and What Might Be Done About It," *Journal of Social Work Education*, vol. 49, no. 1, pp. 136–149, jan 2013. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10437797.2013.755399

[96] C. D. Batson and N. Ahmad, "Empathy induced altruism in a prisoner's dilemma: What if the target of empathy has defected?" *European Journal of Social Psychology*, vol. 36, no. January 2000, pp. 25–36, 2001. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/ejsp.26/full

[97] J. Sierksma, J. Thijs, and M. Verkuyten, "In-group bias in children's intention to help can be overpowered by inducing empathy," *British Journal of Developmental Psychology*, vol. 33, no. 1, pp. 45–56, mar 2015. [Online]. Available: http://doi.wiley.com/10.1111/bjdp.12065

[98] C. D. Batson, J. Chang, R. Orr, and J. Rowland, "Empathy, Attitudes, and Action: Can Feeling for a Member of a Stigmatized Group Motivate One to Help the Group?" *Personality and Social Psychology Bulletin*, vol. 28, no. 12, pp. 1656–1666, dec 2002. [Online]. Available: http://journals.sagepub.com/doi/10.1177/014616702237647

[99] M. Djikic, K. Oatley, and M. C. Moldoveanu, "Reading other minds: Effects of literature on empathy," *Scientific Study of Literature*, vol. 3, no. 1, pp. 28–47, 2013. [Online]. Available: http://www.jbe-platform.com/content/journals/10.1075/ssol.3.1.06dji

[100] T. T. Brunyé, T. Ditman, C. R. Mahoney, J. S. Augustyn, and H. a. Taylor, "When You and I Share Perspectives," *Psychological Science*, vol. 20, no. 1, pp. 27–32, jan 2009. [Online]. Available: http://journals.sagepub.com/doi/10.1111/j.1467-9280.2008.02249.x

[101] C. Lamm, C. D. Batson, and J. Decety, "The Neural Substrate of Human Empathy: Effects of Perspective-taking and Cognitive Appraisal," *Journal of Cognitive Neuroscience*, vol. 19, no. 1, pp. 42–58, jan 2007. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/jocn.2007.19.1.42

[102] M. H. Davis, *Empathy: A Social Psychological Approach*. Westview Press, 1995.

[103] R. L. Reniers, R. Corcoran, R. Drake, N. M. Shryane, and B. A. Völlm, "The QCAE: A questionnaire of cognitive and affective empathy," *Journal of Personality Assessment*, vol. 93, no. 1, pp. 84–95, 2011.

[104] C. D. Batson, M. P. Polycarpou, E. Harmon-Jones, H. J. Imhoff, E. C. Mitchener, L. L. Bednar, T. R. Klein, and L. Highberger, "Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group?" *Journal of Personality and Social Psychology*, vol. 72, no. 1, pp. 105–118, 1997. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.72.1.105

[105] J. D. Foubert and B. C. Perry, "Creating Lasting Attitude and Behavior Change in Fraternity Members and Male Student Athletes," *Violence Against Women*, vol. 13, no. 1, pp. 70–86, jan 2007. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1077801206295125

[106] J. Cohen, "Defining Identification: A Theoretical Look at the Identification of Audiences With Media Characters," *Mass Communication and Society*, vol. 4, no. 3, pp. 245–264, 2001. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/S15327825MCS0403{_}01

[107] K. Oatley, "A taxonomy of the emotions of literary response and a theory of identification in fictional narrative," *Poetics*, vol. 23, no. 1-2, pp. 53–74, 1995.

[108] S. Keen, "A Theory of Narrative Empathy," *Narrative*, vol. 14, no. 3, pp. 207–236, 2006. [Online]. Available: http://muse.jhu.edu/content/crossref/journals/narrative/v014/14.3keen.html

[109] S. Stürmer, M. Snyder, A. Kropp, and B. Siem, "Empathy-Motivated Helping: The Moderating Role of Group Membership," *Personality and Social Psychology Bulletin*, vol. 32, no. 7, pp. 943–956, jul 2006. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0146167206287363

[110] G. Hein, G. Silani, K. Preuschoff, C. D. Batson, and T. Singer, "Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping," *Neuron*, vol. 68, no. 1, pp. 149–160, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2010.09.003

[111] J. N. Gutsell and M. Inzlicht, "Intergroup differences in the sharing of emotive states: Neural evidence of an empathy gap," *Social Cognitive and Affective Neuroscience*, vol. 7, no. 5, pp. 596–603, 2012.

[112] M. Cikara, E. G. Bruneau, and R. R. Saxe, "Us and Them: Intergroup Failures of Empathy," *Current Directions in Psychological Science*, vol. 20, no. 3, pp. 149–153, jun 2011. [Online]. Available: http://cdp.sagepub.com/lookup/doi/10.1177/0963721411408713http://journals.sagepub.com/doi/10.1177/0963721411408713

[113] M. Tarrant, S. Dazeley, and T. Cottom, "Social categorization and empathy for outgroup members," *British Journal of Social Psychology*, vol. 48, no. 3, pp. 427–446, sep 2009. [Online]. Available: http://doi.wiley.com/10.1348/014466608X373589

[114] W. G. Stephan and K. Finlay, "The Role of Empathy in Improving Intergroup Relations," *Journal of Social Issues*, vol. 55, no. 4, pp. 729–743, jan 1999. [Online]. Available: http://doi.wiley.com/10.1111/0022-4537.00144

[115] C. Nass, B. Fogg, and Y. Moon, "Can computers be teammates?" *International Journal of Human-Computer Studies*, vol. 45, no. 6, pp. 669–678, dec 1996. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1071581996900737

[116] B. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani et al., "What makes web sites credible? a report on a large quantitative study," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2001, pp. 61–68.

[117] D. Running, J. Ligon, and I. Miskioglu, "Better Liked Than Right: Trustworthiness and Expertise as Factors in Credibility," *Journal of Composite Materials*, vol. 33, no. 10, pp. 928–940, 1999.

[118] Q. V. Liao, W.-t. Fu, N. G. Avenue, and U. Il, "Expert Voices in Echo Chambers : Effects of Source Expertise Indicators on Exposure to Diverse Opinions," in *CHI*, 2014.

[119] C. T. Carr and J. B. Walther, "Increasing attributional certainty via social media: Learning about others one bit at a time," *Journal of Computer-Mediated Communication*, vol. 19, no. 4, pp. 922–937, 2014.

[120] D. Kanouse, "Explaining Negativity Biases in Evaluation and Choice Behavior: Theory and Research," *Advances in Consumer Research*, vol. 11, pp. 703–708, 1984.

[121] J. Marlow and L. A. Dabbish, "The Effects of Visualizing Activity History on Attitudes and Behaviors in a Peer Production Context," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. New York, New York, USA: ACM Press, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2675133.2675250 pp. 757–764.

[122] P. Siangliulue, K. C. Arnold, K. Z. Gajos, and S. P. Dow, "Toward Collaborative Ideation at Scale," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. New York, New York, USA: ACM Press, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2675133.2675239 pp. 937–945.

[123] S.-w. Huang and W.-t. Fu, "Don ' t Hide in the Crowd ! Increasing Social Transparency Between Peer Workers Improves Crowdsourcing Outcomes," in *CHI*, 2013, pp. 621–630.

[124] J. Cheng, J. Teevan, and M. S. Bernstein, "Measuring Crowdsourcing Effort with Error-Time Curves," in *CHI*, 2015.

[125] Y. Gao and A. Parameswaran, "Finish them!" *Proceedings of the VLDB Endowment*, vol. 7, no. 14, pp. 1965–1976, oct 2014. [Online]. Available: http://dl.acm.org/doi/10.14778/2733085.2733101

[126] V. Ambati, S. Vogel, and J. Carbonell, "Towards Task Recommendation in Micro-Task Markets." in *AAAI*, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewFile/4005/4264

[127] W. Willett, J. Heer, and M. Agrawala, "Strategies for crowdsourcing social data analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 227–236.

[128] K. Luther, K. Caine, K. Ziegler, and A. Bruckman, "Why It Works (When It Works): Success Factors in Online Creative Collaboration," in *GROUP*, vol. 10, 2010, pp. 1–10.

[129] M. J. Metzger, A. J. Flanagin, and R. B. Medders, "Social and heuristic approaches to credibility evaluation online," *Journal of Communication*, vol. 60, no. 3, pp. 413–439, 2010.

[130] G. Peeters and J. Czapinski, "Positive-Negative Asymmetry in Evaluations: The Distinction Between Affective and Informational Negativity Effects," *European Review of Social Psychology*, vol. 1, no. 1, pp. 33–60, jan 1990. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14792779108401856

[131] R. P. Bagozzi and D. J. Moore, "Public Service Advertisements: Emotions and Empathy Guide Prosocial Behavior," *Journal of Marketing*, vol. 58, no. 1, p. 56, jan 1994. [Online]. Available: http://www.jstor.org/stable/1252251?origin=crossrefhttps://www.jstor.org/stable/1252251?origin=crossref

[132] K. H. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2004.

[133] M. Bortolussi and P. Dixon, *Psychonarratology*. Cambridge: Cambridge University Press, 2002. [Online]. Available: https://www.cambridge.org/core/product/identifier/9780511500107/type/book

[134] J. Shanteau, D. J. Weiss, R. P. Thomas, and J. C. Pounds, "Performance-based assessment of expertise: How to decide if someone is an expert or not," *European Journal of Operational Research*, vol. 136, no. 2, pp. 253–263, 2002.

[135] S. Dow, J. Fortuna, D. Schwartz, B. Altringer, D. Schwartz, and S. Klemmer, "Prototyping dynamics," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. New York, New York, USA: ACM Press, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1978942.1979359 p. 2807.

[136] T. Mitra, C. J. Hutto, and E. Gilbert, "Comparing person- and process-centric strategies for obtaining quality data on Amazon Mechanical Turk," in *CHI*. New York, New York, USA: ACM Press, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2702123.2702553 pp. 1345–1354.

[137] D. Prelec, "A Bayesian truth serum for subjective data." *Science (New York, N.Y.)*, vol. 306, no. 5695, pp. 462–6, oct 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15486294

[138] K. Z. Gajos, K. Reinecke, and C. Herrmann, "Accurate Measurements of Pointing Performance from In Situ Observations," *CHI*, no. Session: Human performance gives us Fitts', pp. 3157–3166, 2012.

[139] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut, "Crowdforge: Crowdsourcing complex work," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 43–52.

[140] S. E. Hudson, J. Fogarty, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, "Predicting Human Interruptibility with Sensors : A Wizard of Oz Feasibility Study," in *CHI*, no. 5, 2003, pp. 257–264.

[141] A. Ghazarian and S. M. Noorhosseini, "Automatic detection of users' skill levels using high-frequency user interface events," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 109–146, jun 2010. [Online]. Available: http://link.springer.com/10.1007/s11257-010-9073-5

[142] A. Hurst, S. E. Hudson, and J. Mankoff, "Dynamic detection of novice vs. skilled use without a task model," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. New York, New York, USA: ACM Press, 2007. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1240624.1240669 pp. 271–280.

[143] P. Salmon, "Effects of physical exercise on anxiety, depression, and sensitivity to stress," *Clinical Psychology Review*, vol. 21, no. 1, pp. 33–61, feb 2001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S027273589900032X

[144] B. L. Wisner, B. Jones, and D. Gwin, "School-based Meditation Practices for Adolescents: A Resource for Strengthening Self-Regulation, Emotional Coping, and Self-Esteem," *Children & Schools*, vol. 32, no. 3, pp. 150–159, jul 2010. [Online]. Available: https://academic.oup.com/cs/article-lookup/doi/10.1093/cs/32.3.150

[145] R. D. Schroeder and J. F. Frana, "SPIRITUALITY AND RELIGION, EMOTIONAL COPING, AND CRIMINAL DESISTANCE: A QUALITATIVE STUDY OF MEN UNDERGOING CHANGE," *Sociological Spectrum*, vol. 29, no. 6, pp. 718–741, oct 2009. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02732170903189076

[146] L. Lu and R. Bol, "A Comparison of Anonymous Versus Identifiable e-Peer Review on College Student Writing Performance and the Extent of Critical Feedback," *Journal of Interactive Online Learning*, vol. 6, no. 2, pp. 100–115, 2007. [Online]. Available: www.ncolr.org/jiol

[147] S. W. McQuiggan and J. C. Lester, "Learning empathy," in *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems - AAMAS '06*. New York, New York, USA: ACM Press, 2006. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1160633.1160806 p. 961.

[148] J. P. Baker and H. Berenbaum, "Emotional approach and problem-focused coping: A comparison of potentially adaptive strategies," *Cognition & Emotion*, vol. 21, no. 1, pp. 95–118, 2007. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02699930600562276

[149] M. M. Nelson and C. D. Schunn, "The nature of feedback: How different types of peer feedback affect writing performance," *Instructional Science*, vol. 37, no. 4, pp. 375–401, 2009.

[150] D. Watson, L. a. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales." *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063–70, 1988. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3397865

[151] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," pp. 707–710, 1966.

[152] L. J. Williams and H. Abdi, "Fisher's least significant difference (lsd) test," *Encyclopedia of research design*, vol. 218, pp. 840–853, 2010.

[153] M. J. Bietz, "Effects of communication media on the interpretation of critical feedback," *Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08*, pp. 467–476, 2008. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-77950791151{\&}partnerID=40{\&}md5=1346673d7492b61abff98c0cfc6530c5http://portal.acm.org/citation.cfm?doid=1460563.1460637

[154] Y.-C. G. Yen, S. P. Dow, E. Gerber, and B. P. Bailey, "Listen to Others, Listen to Yourself," in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition - C&C '17*. New York, New York, USA: ACM Press, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3059454.3059468 pp. 158–170.

[155] J. Falout, J. Elwood, and M. Hood, "Demotivation: Affective states and learning outcomes," *System*, vol. 37, no. 3, pp. 403–417, sep 2009. [Online]. Available: http://dx.doi.org/10.1016/j.system.2009.03.004http://linkinghub.elsevier.com/retrieve/pii/S0346251X09000566

[156] P. Cantillon and J. Sargeant, "Giving feedback in clinical settings." *BMJ (Clinical research ed.)*, vol. 337, no. 7681, p. a1961, 2008.

[157] C. M. Neuwirth, R. Chandhok, D. Charney, P. Wojahn, and L. Kim, "Distributed collaborative writing," in *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. New York, New York, USA: ACM Press, 1994. [Online]. Available: http://portal.acm.org/citation.cfm?doid=191666.191693 pp. 51–57.

[158] D. B. Willingham, "Effective feedback on written assignments," *Teaching of Psychology*, vol. 17, no. 1, pp. 10–13, 1990. [Online]. Available: https://login.proxy.library.emory.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true{\&}db=tfh{\&}AN=6390195{\&}site=ehost-live

[159] J. Shah, "Hacker News: Stop greying out flagged comments (March 10, 2012)," 2012. [Online]. Available: http://blog.jasonshah.org/post/19079124354/hacker-news-stop-greying-out-flagged-comments

[160] L. Hale, "What is the meaning of "answer collapsed" in Quora? (August 22, 2015)," 2015. [Online]. Available: https://www.quora.com/What-is-the-meaning-of-answer-collapsed-in-Quora/answer/Laura-Hale?srid=uxyk

[161] D. R. Ferris, "The Influence of Teacher Commentary on Student Revision," *TESOL Quarterly*, vol. 31, no. 2, p. 315, 1997. [Online]. Available: http://www.jstor.org/stable/3588049?origin=crossref

[162] C. Lampe, R. Wash, A. Velasquez, and E. Ozkaya, "Motivations to participate in online communities," in *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. New York, New York, USA: ACM Press, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1753326.1753616 p. 1927.

[163] "Sentiment Analysis API— AlchemyAPI," p. http://www.alchemyapi.com/.

[164] S. M. Smith, J. S. Linsey, and A. Kerne, "Using Evolved Analogies to Overcome Creative Design Fixation," in *Design Creativity 2010*. London: Springer London, 2011, pp. 35–39. [Online]. Available: http://link.springer.com/10.1007/978-0-85729-224-7{_}6

[165] J. van der Pol, B. van den Berg, W. Admiraal, and P. Simons, "The nature, reception, and use of online peer feedback in higher education," *Computers & Education*, vol. 51, no. 4, pp. 1804–1817, dec 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0360131508000833

[166] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?" in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*. New York, New York, USA: ACM Press, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1753846.1753873 p. 2863.

[167] A. A. Nease, B. O. Mudgett, and M. A. Quiñones, "Relationships among feedback sign, self-efficacy, and acceptance of performance feedback." *Journal of Applied Psychology*, vol. 84, no. 5, pp. 806–814, 1999. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.84.5.806

[168] T.-A. Roberts and S. Nolen-Hoeksema, "Sex differences in reactions to evaluative feedback," *Sex Roles*, vol. 21, no. 11-12, pp. 725–747, dec 1989. [Online]. Available: http://link.springer.com/10.1007/BF00289805

[169] S. R. Finkelstein and A. Fishbach, "Tell Me What I Did Wrong: Experts Seek and Respond to Negative Feedback," *Journal of Consumer Research*, vol. 39, no. 1, pp. 22–38, jun 2012. [Online]. Available: http://jcr.oxfordjournals.org/lookup/doi/10.1086/661934

[170] C. Happ, A. Melzer, and G. Steffgen, "Bringing Empathy into Play: On the Effects of Empathy in Violent and Nonviolent Video Games," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6972 LNCS, pp. 371–374. [Online]. Available: http://link.springer.com/10.1007/978-3-642-24500-8{_}44

[171] W. G. Graziano, M. M. Habashi, B. E. Sheese, and R. M. Tobin, "Agreeableness, empathy, and helping: A person × situation perspective." *Journal of Personality and Social Psychology*, vol. 93, no. 4, pp. 583–599, 2007. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.93.4.583

[172] M. J. Traxler and M. A. Gernsbacher, "Improving written communication through perspective-taking," *Language and Cognitive Processes*, vol. 8, no. 3, pp. 311–334, aug 1993. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01690969308406958

[173] P. A. Oswald, "The Effects of Cognitive and Affective Perspective Taking on Empathic Concern and Altruistic Helping," *The Journal of Social Psychology*, vol. 136, no. 5, pp. 613–623, oct 1996. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00224545.1996.9714045

[174] D. Weller and K. Hansen Lagattuta, "Helping the In-Group Feels Better: Children's Judgments and Emotion Attributions in Response to Prosocial Dilemmas," *Child Development*, vol. 84, no. 1, pp. 253–268, jan 2013. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-8624.2012.01837.x

[175] E. A. Smith and J. P. Kincaid, "Derivation and Validation of the Automated Readability Index for Use with Technical Materials," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 12, no. 5, pp. 457–564, oct 1970. [Online]. Available: http://journals.sagepub.com/doi/10.1177/001872087001200505

[176] T. Hamby and W. Taylor, "Survey Satisficing Inflates Reliability and Validity Measures," *Educational and Psychological Measurement*, vol. 76, no. 6, pp. 912–932, dec 2016. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0013164415627349

[177] A. Xu, H. Rao, S. P. Dow, and B. P. Bailey, "A Classroom Study of Using Crowd Feedback in the Iterative Design Process," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. New York, New York, USA: ACM Press, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2675133.2675140 pp. 1637–1648.

[178] X. Ma, L. Yu, J. L. Forlizzi, and S. P. Dow, "Exiting the design studio: Leveraging online participants for early-stage design feedback," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015, pp. 676–685.

[179] H. Wauck, Y.-C. G. Yen, W.-T. Fu, E. Gerber, S. P. Dow, and B. P. Bailey, "From in the Class or in the Wild?" in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, vol. 2017-May. New York, New York, USA: ACM Press, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3025453.3025477 pp. 5580–5591.

[180] S. P. Dow, E. M. Gerber, and A. Wong, "A pilot study of using crowds in the classroom," *CHI*, p. 227, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2470654.2470686

[181] T. J. Ngoon, C. A. Fraser, A. S. Weingarten, M. Dontcheva, and S. Klemmer, "Interactive Guidance Techniques for Improving Creative Feedback," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. New York, New York, USA: ACM Press, 2018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3173574.3173629 pp. 1–11.

[182] J. Seering, T. Fang, L. Damasco, M. C. Chen, L. Sun, and G. Kaufman, "Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. New York, New York, USA: ACM Press, 2019. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3290605.3300836 pp. 1–14.

[183] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, mar 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0261927X09351676

[184] A. B. Nickerson, A. M. Aloe, J. A. Livingston, and T. H. Feeley, "Measurement of the bystander intervention model for bullying and sexual harassment," *Journal of Adolescence*, vol. 37, no. 4, pp. 391–400, jun 2014. [Online]. Available: http://dx.doi.org/10.1016/j.adolescence.2014.03.003https://linkinghub.elsevier.com/retrieve/pii/S0140197114000323

[185] M. H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach." *Journal of Personality and Social Psychology*, vol. 44, no. 1, pp. 113–126, 1983. [Online]. Available: http://content.apa.org/journals/psp/44/1/113

[186] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of Other in the Self Scale and the structure of interpersonal closeness." *Journal of Personality and Social Psychology*, vol. 63, no. 4, pp. 596–612, 1992. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.63.4.596

[187] P. J. Loewen, G. Lyle, and J. S. Nachshen, "An eight-item form of the Empathy Quotient (EQ) and an application to charitable giving," pp. 1–14, 2010.

[188] H. Abdi, "Holm's sequential bonferroni procedure," *Encyclopedia of research design*, vol. 1, no. 8, pp. 1–8, 2010.

[189] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," *Austin, TX: University of Texas at Austin*, pp. 1–22, 2015. [Online]. Available: http://www.liwc.net/LIWC2007LanguageManual.pdf

[190] J. Antin and A. Shaw, "Social desirability bias and self-reports of motivation," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. New York, New York, USA: ACM Press, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2208699http://dl.acm.org/citation.cfm?doid=2207676.2208699 p. 2925.

[191] A. Richterich, "' Karma , Precious Karma !' Karmawhoring on Reddit and the Front Page ' s Econometrisation," *Journal of Peer Production*, no. May, pp. 1–12, 2013. [Online]. Available: http://peerproduction.net

[192] D. Retelny, S. Robaszkiewicz, A. To, W. S. Lasecki, J. Patel, N. Rahmati, T. Doshi, M. Valentine, and M. S. Bernstein, "Expert crowdsourcing with flash teams," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 75–85.

[193] J. S. Hui, D. Gergle, and E. M. Gerber, "IntroAssist," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. New York, New York, USA: ACM Press, 2018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3173574.3173596 pp. 1–13.

[194] C. A. Fraser, J. O. Kim, H. V. Shin, J. Brandt, and M. Dontcheva, "Temporal Segmentation of Creative Live Streams," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, apr 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376437 pp. 1–12.

[195] Z. Liu, Z. Liu, and T. Munzner, "Data-driven Multi-level Segmentation of Image Editing Logs," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, apr 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376152 pp. 1–12.

[196] G. Krampen, "Promotion of creativity (divergent productions) and convergent productions by systematic-relaxation exercises: empirical evidence from five experimental studies with children, young adults, and elderly," *European Journal of Personality*, vol. 11, no. 2, pp. 83–99, jun 1997. [Online]. Available: http://doi.wiley.com/10.1002/{\%}28SICI{\%}291099-0984{\%}28199706{\%}2911{\%}3A2{\%}3C83{\%}3A{\%}3AAID-PER280{\%}3E3.0.CO{\%}3B2-5

[197] C. J. Brown, A. C. N. Chen, and S. F. Dworkin, "Music in the Control of Human Pain," *Music Therapy*, vol. 8, no. 1, pp. 47–60, jan 1989. [Online]. Available: https://academic.oup.com/musictherapy/article-lookup/doi/10.1093/mt/8.1.47

[198] R. Deloatch, B. P. Bailey, A. Kirlik, and C. Zilles, "I Need Your Encouragement! Requesting Supportive Comments on Social Media Reduces Test Anxiety," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2017.

[199] N. David, B. H. Bewernick, M. X. Cohen, A. Newen, S. Lux, G. R. Fink, N. J. Shah, and K. Vogeley, "Neural Representations of Self versus Other: Visual-Spatial Perspective Taking and Agency in a Virtual Ball-tossing Game," *Journal of Cognitive Neuroscience*, vol. 18, no. 6, pp. 898–910, jun 2006. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/jocn.2006.18.6.898

[200] M. J. Chandler, "Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills." *Developmental Psychology*, vol. 9, no. 3, pp. 326–332, 1973. [Online]. Available: http://content.apa.org/journals/dev/9/3/326

[201] D. Zillmann, "Mechanisms of emotional involvement with drama," *Poetics*, vol. 23, no. 1-2, pp. 33–51, jan 1995. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0304422X94000207

[202] S. Rafaeli, Quentin Jones, "Time to Split, Virtually: 'Discourse Architecture' and 'Community Building' Create Vibrant Virtual Publics," *Electronic Markets*, vol. 10, no. 4, pp. 214–223, oct 2000. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/101967800750050326

[203] C. Ruggles, G. Wadley, and M. R. Gibbs, "Online Community Building Techniques Used by Video Game Developers," 2005, pp. 114–125. [Online]. Available: http://link.springer.com/10.1007/11558651{_}12

[204] D. Dannels, A. Gaffney, and K. Martin, "Beyond Content, Deeper than Delivery: What Critique Feedback Reveals about Communication Expectations in Design Education," *International Journal for the Scholarship of Teaching and Learning*, vol. 2, no. 2, jan 2008. [Online]. Available: http://digitalcommons.georgiasouthern.edu/ij-sotl/vol2/iss2/12